AD-A252 401

...TION PAGE

OMB No. 0704-0188

...verage 1 hour per response, including the time for reviewing instructions, searching existing data sources, ...g the collection of information. Send comments regarding this burden estimate or any other aspect of this ...o Washington Headquarters Services, Directorate for information Operations and Reports, 1215 Jefferson ...Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503

...TE	3. REPORT TYPE AND DATES COVERED
19-May-92	Final 1-Jun-91 to 19-May-92

4. TITLE AND SUBTITLE
Optimization of Resonant Interband Tunnel Devices

5. FUNDING NUMBERS
F49620-91-C-0043

(2)

6. AUTHOR(S)
Dr. Mark F. Sweeny

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Zytron Ltd.
85 N Cretin Av.
St. Paul, MN 55104

8. PERFORMING ORGANIZATION REPORT NUMBER
AF-1

AFOSR-TR- [illegible]

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
USAF, AFSC
Air Force Office Of Scientific Research
Building 410
Bolling AFB DC 20332-6448

NE

10. SPONSORING / MONITORING AGENCY REPORT NUMBER
3005/A1

11. SUPPLEMENTARY NOTES

DTIC
S JUN 22 1992 D
A

12a. DISTRIBUTION / AVAILABILITY STATEMENT
unlimited

12b. DISTRIBUTION CODE

This document has been approved for public release and sale; its distribution is unlimited.

13. ABSTRACT *(Maximum 200 words)*

We have carried out analytical and numerical studies of resonant interband tunnelling. The numerical methods implement a 2 band model, consisting of the conduction band and a single valence band. The valence band can be considered to be the "light" holes. The numerical methods are described in detail, and can be applied to multiband models too. Analytic estimates are made of the thermionic currents, the of effect of stress, and other physical effects not included in the numerical models. The devices simulated are closely modeled after a set of Resonant Interband Tunnel Diodes fabricated at the Varian Research Center in Palo Alto California. Comparison of the measured and computed results show that our simulator predicts the maximal currents to within a factor of two, for devices with maximal currents varying by a factor of 1000. There are systematic differences which are likely to be due to the very high doping used in the devices. Finally, we describe a graphic user interface, implemented for our device simulator, and a *Mathematica* package for carrying out symbolic computations upon the operators of the Luttinger model and related multiband models of semiconductors.

14. SUBJECT TERMS
Semiconductor, Tunneling, Device Modeling
Resonant Interband Tunneling

15. NUMBER OF PAGES
46

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
...IED	UNCLASSIFIED	[illegible]	UL

NSN 7540-01-280-5500

Standard Form 298 (Rev 2-89)
Prescribed by ANSI Std Z39-18
298-102

Optimization of Resonant Interband Tunnel Devices

Mark F Sweeny

Zytron Ltd

Abstract

We have carried out analytical and numerical studies of resonant interband tunnelling. The numerical methods implement a 2 band model, consisting of the conduction band and a single valence band. The valence band can be considered to be the "light" holes. The numerical methods are described in detail, and can be applied to multiband models too. Analytic estimates are made of the thermionic currents, the of effect of stress, and other physical effects not included in the numerical models. The devices simulated are closely modeled after a set of Resonant Interband Tunnel Diodes fabricated at the Varian Research Center in Palo Alto California. Comparison of the measured and computed results show that our simulator predicts the maximal currents to within a factor of two, for devices with maximal currents varying by a factor of 1000. There are systematic differences which are likely to be due to the very high doping used in the devices. Finally, we describe a graphic user interface, implemented for our device simulator, and a *Mathematica* package for carrying out symbolic computations upon the operators of the Luttinger model and related multiband models of semiconductors. The work described here was funded by a contract with the US Air Force Office of Scientific Research

Introduction

In this report, we describe a project to carry out numerical and analytic studies of RITD's, which was funded by a contract with the US Air Force Office of Scientific Research

Resonant Interband Tunnel Diodes (RITD's) were proposed independently by several researchers[1, 2, 3]. These devices are essentially tunnel diodes[4]. Which incorporate heterostructures. Such devices have achieved peak to valley current ratio's in excess of 100 at room temperature[5] and current densities in excess of $10^9 A/m^2$ have been observed[6], making these devices promising for use in high speed electronics. At this time the publications relating to RITD's, are so numerous that it is impossible to give a treatment to all authors. I would point out that McGill et. al. probably have carried out the most sophisticated device simulations, to date[7]. Their work has focused upon the GaSb/InAs/AlSb material system, while the present study uses material parameters for InAlAs and InGaAs lattice matched to InP. In this report, we compare our numerical results to the experimental results obtained for a set of devices made in this material system at the Varian Research Center, in Palo Alto California[8, 9].

While this was not, strictly speaking, a software project, the bulk of the time spent on the project actually consisted of software development. Because of this, and because the software developed has potential both for future research and for development into commercial products, documentation of the software is a major part of this report.

The sections on the basic device physics and on the theory of interband tunnelling are meant to provide background. This material has in its essential form been published, and is too complex to be dealt with fairly in a report of this length. Part of the function of such an expository section is simply to define the notation. It is also hoped, however, that a reader with a strong background in general physics should be able to gain some understanding of the subject.

As a part of the research carried out under this contract, we implemented a tensor package to manipulate the algebraic objects arising in the theory of the valence bands of the zincblend semiconductors. The section describing this work is the most mathematically demanding part of the report. The material in that section, is, however not needed for any other parts of the report, and so can be skipped if need be.

RTD Device Physics

At a simplified level, Resonant Interband Tunnel Diodes (RITD's) may be understood by reference to Figure 1. This figure illustrates Interband tunnelling, thermionic emission, and indirect conduction by means of scattering into intermediate bound states and tunnelling across the small barrier. The case 1c is used to illustrate the fact that in our paradigm the conduction through a device with very wide wells is dependent upon scattering to populate bound states within the wells. This complex explanation contrasts with the physically intuitive concept that the device in 1c is not a double-well device at all, but a single barrier device, with contacts to high-band gap electrodes "located far from the region of interest." Of course, computerized device simulators are not very intelligent, and it is important to note that the thermionic emission current in a device such as that shown in Figure 1 is likely to be underestimated if "direct" currents only are calculated.

Of course, a numerical device simulation must solve the Poisson's equation self-consistently, with the carrier density, but this problem will be treated later, as it does not involve any new physics that is special to RITD's.

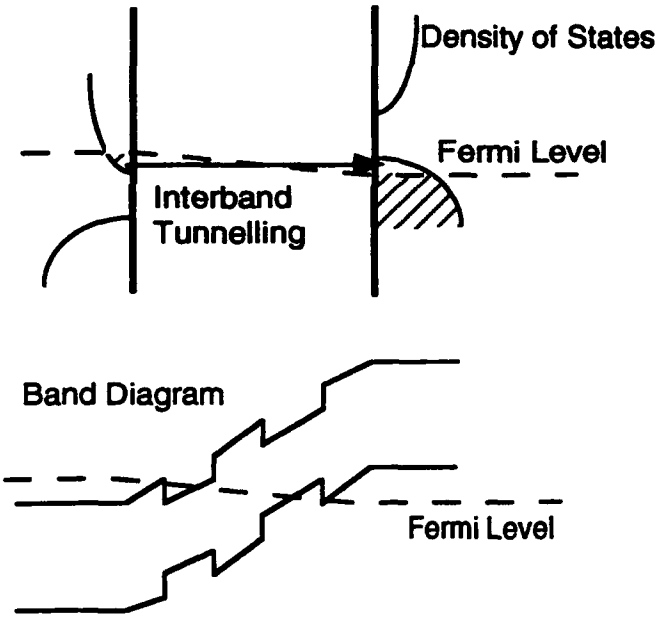


Fig 1a Interband Tunneling

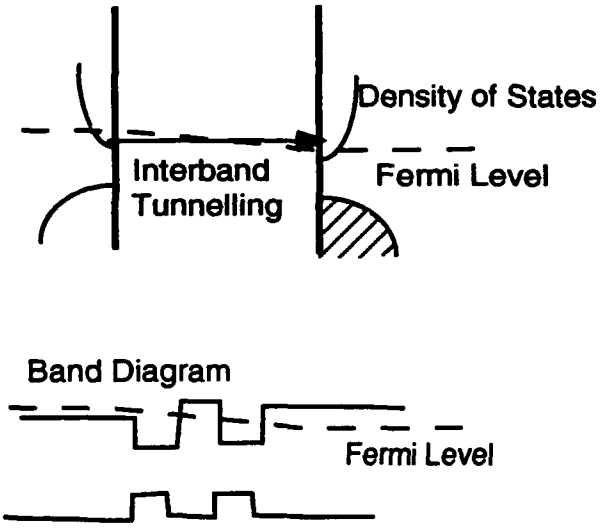


Fig 1b Thermionic Emmission

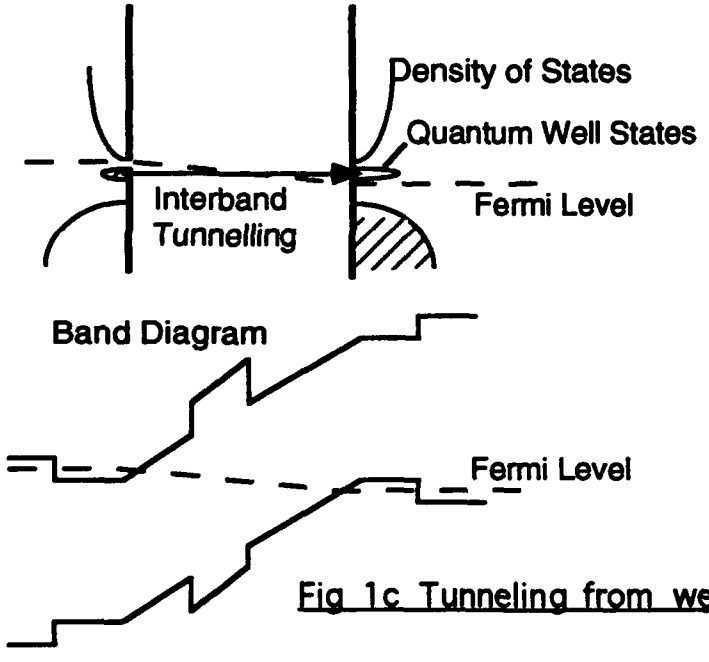


Fig 1c Tunneling from well states

This schematically illustrates tunnelling in an RITD, showing Interband tunnelling, thermionic emission, and indirect conduction by means of scattering into intermediate bound states. This project has primarily considered direct interband tunnelling. In 1a, the device is biased with a small voltage, and is highly conductive by means of interband tunnelling. In Figure 1b, the forward bias is large enough to cause a heavy thermionic emission current to flow. Fig 1 c illustrates the case of an RITD with two very wide quantum wells. In this case the conduction by means of scattering is important.

Figure 1

Theory of RITD Devices

The theory of RTD devices is based upon the band structure indicated in Figure 2, which is a generic structure typical of zincblend and diamond semiconductors at the Γ point of the band. This band structure consists of a conduction band and a set of valence bands. It is interesting to put this into historical perspective. Shockley[10] is (as far as this author knows) the first to suggest that the valence bands would be degenerate (because they are transformed according to an irreducible representation of the cubic group) at the Γ point. His statement, which is essentially correct, was that the valence bands form a vector and that the wave equation is analogous to sound in an elastic medium, with different velocities for longitudinal (compressive) waves and for transverse (shear) waves. A few years later, when Dresselhaus, Kip, and Kittel[11] (in an exceptionally good paper) described and interpreted cyclotron resonance experiments, it became clear that Shockley's model neglected the spin orbit splitting. Schockley thought of the valence bands as a vector, with the spin 1/2 nature of the electron playing no essential role. Kip, Kittel and Dresselhaus found that the spin 1 degree of freedom inherent in the vector nature of the band mixes with the spin 1/2 electron spin via the spin orbit interaction. While the spin orbit interaction is normally considered "weak" in this case, the original bands are degenerate, and so the splitting between the "light and heavy" holes at the top of the band and the split-off band below is significant. Cyclotron resonance experiments concern a situation of very low temperature, and relatively low doping, where even the split-off band can be considered "far" from the band maximum. In this case, one can write a separate Hamiltonian for just the spin 3/2 components, which has come to be known as the Luttinger Hamiltonian. Unfortunately, this Hamiltonian involves 4 by 4 spin 3/2 matrices, and this has caused the field to suffer from algebraic complexity ever since.

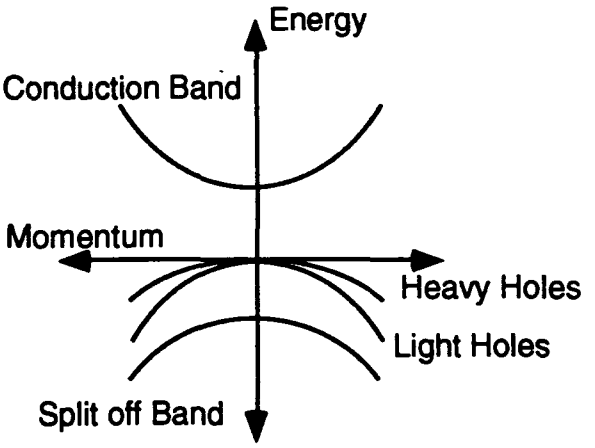


Figure 2 Band Structure

The Band Structure of a "typical" zincblend semiconductor at the Γ point. The conduction band and split-off band are both two-fold degenerate, due to spin, while the light and heavy holes are fourfold degenerate at the the Γ point. As k increases, the light and heavy holes split apart, as shown. It is important to realize that the light and heavy holes are really different polarizations of the same "particle," so that scattering, for example, will generally cause a mixing.

As if a complex valence band structure were not enough, Kane[12], in explaining the band structure of Indium Antimonide, reasoned that the conduction band was too close to the valence bands for any interactions to be treated as small perturbations, and so arrived at the well-known Kane Hamiltonian, which couples all eight bands at once. The spherical Kane Hamiltonian is an 8-by-8 matrix equation, but can be written

$$\mathbf{H}\begin{bmatrix}\theta\\ \vec{\Psi}\end{bmatrix}=\begin{bmatrix}A_c(-\nabla^2)+V_c(r)+E_g & -P\nabla\cdot\\ P\vec{\nabla} & A_v\nabla^2+B\vec{\nabla}\nabla\cdot-\frac{\Delta}{3}\vec{\sigma}\sigma\cdot-V_v(r)\end{bmatrix}\begin{bmatrix}\theta\\ \vec{\Psi}\end{bmatrix} \qquad [1]$$

where θ is a spin 1/2 envelope function representing the conduction band, and $\vec{\Psi}$ is a set of three spin 1/2 envelope functions,

$$\vec{\Psi} = (\psi_x, \psi_y, \psi_z)$$

$\vec{\Psi}$ is a tensor product between a vector and a spinor representation of SU(2), and is itself a reducible representation, having a spin 3/2 component, and a spin 1/2 component. The spin 3/2 components become the light and heavy holes, while the spin 1/2 components become the split-off band. Note that the second column contains vector dot products with $\vec{\Psi}$, and that while ∇ and σ are generally vectors, a vector symbol has been used selectively to emphasize the vector whose index appears on the left side of the equation. The tensor product approach, which is a favorite of this author, makes the simplicity of Shockley's first model become apparent, but (as can be verified with algebra) requires us to note that the spin orbit interaction can be written very simply in terms of a vector dot product with σ matrices operating upon the spin 1/2 degrees of freedom

$$\Delta(L \cdot \sigma - 1) = -\frac{\Delta}{3}\vec{\sigma}\sigma \cdot ,$$

where the -1 on the left is to make the term be zero for the spin 3/2 component, and $-\Delta$ for the split-off spin 1/2 component. It must also be stated that the full Kane Hamiltonian (as well as the Luttinger) contain "band warping" terms. These are additional anisotropic interactions, which break spherical (but not cubic) symmetry. While the band warping terms are responsible for much of the complexity of the theory, they are also numerically small compared to the other terms in the Hamiltonian. They are primarily important in the dispersion of the heavy holes.

In this work, we begin with a two-band model, loosely modeled after the Kane Hamiltonian,

$$\mathbf{H}\begin{bmatrix}\theta \\ \Psi\end{bmatrix} = \begin{bmatrix} A_c(-\nabla^2) + V_c(r) + E_g & -P\nabla \\ P\nabla & A_v\nabla^2 - V_v(r)\end{bmatrix}\begin{bmatrix}\theta \\ \Psi\end{bmatrix} \qquad [2]$$

This Hamiltonian was used by Kane[13] in his own modelling of interband tunnelling in homojunction tunnel diodes. The similarity to equation 1 is less than it appears, as Ψ is no longer a vector, and the spin orbit interaction is now being ignored. This Hamiltonian has the obvious advantage of (relative) simplicity. As interband tunnelling requires at least two bands, which must be coupled, this is the minimal Hamiltonian for the modelling of such phenomena. While it is easy to accept the proposition that the 2-band model accounts qualitatively for the physics of interband tunnelling, it takes further convincing to believe that it should be quantitatively close to the 8-band model. We will return to this point in the following section.

Band Structure and Tunnelling

Physically, the dispersion relation or momentum (k) verses energy (E) is the primary characteristic of a band, and so it is instructive to consider the dispersion relations for the 8-band and 2-band models. The dispersion relation is equivalent to solving the secular equation for the Hamiltonian, with ∇ replaced by ik, the momentum of a plane wave.

$$\text{Determinant}[\mathrm{H}(k) - E\ \mathrm{I}] = 0$$

In his famous paper on the band structure of InSb, Kane elucidated the band structure resulting from the 8-band Hamiltonian. To a very good approximation, the heavy holes form a doubly degenerate band, which does not interact at all with the other bands. The remaining six bands are paired into 3 doubly degenerate bands. The double degeneracy of all bands means that the secular equation, which is an 8th-order polynomial in E, is actually a square of a quartic. The fact that the heavy holes do not interact with the other bands means that heavy hole solution can be separated out, and we are left with a cubic for E:

$$E'(E'-E_G)(E'+\Delta)-k^2P^2\left(E'+\frac{2\Delta}{3}\right)=0$$

,

where $E'=E-\frac{\hbar}{2m_e}k^2$, the energy minus the bare electron energy.

The three solutions for E represent the conduction band, the light holes, and the split-off band. An illustration of this band structure was shown earlier as Figure 2.

In deriving this equation, Kane assumed that the diagonal entries in H were of the bare-electron form, but when one allows for arbitrary values, a cubic still results, although it is more complex. It is this more complex cubic which has been used at ZYTRON, and forms the basis for the numerical results that make up the remainder of this section.

Figure 3 shows the computed band structures resulting from these equations. Numerically, they are very close, and one would not expect that there would be a big difference in tunneling rates due to choice of band model. For InAlAs, the barrier is slightly deeper for the case of the 2-band model, and so the overall tunnel rate would be expected to be less. Because tunnel current is exponential in the barrier height, the difference is $|T_8|^2 \approx \left(|T_2|^2\right)^{(0.82)}$, where T_8 is for the 8-band tunnel coefficient, and T_2 is for the 2-band model.

It is worth noticing that the dispersion curves tend to be very close (numerically) to the following simple formula:

$$\frac{\hbar^2}{2m}k^2=\frac{(E-E_c)(E+E_v)}{E_g}$$

In fact this curve lies (numerically) between the two shown in Figure 3. With this formula one can perform the WKB integral exactly to estimate tunnelling through the forbidden region in the case of a traditional tunnel diode:

$$\log(T)\approx-\frac{L}{E_g}\int_{E_v}^{E_c}dE\ \mathrm{Im}(k)=\frac{\pi}{8}L\sqrt{\frac{2m}{\hbar^2}}\,E_g^{3/2}$$

where L is the length, of the forbidden region, and the potential is assumed to vary linearly from E_V to E_C as one traverses the forbidden region.

Looking again at Figure 3, the points where the curve intersects the energy axis are fixed by the band gap, and the slopes at those points are fixed by the physical masses. This heavily constrains the shape of the curve, with the result that the overall depends very weakly upon the matrix element P, provided of course that P is not 0 and that the physical masses are held constant by the adjustment of the other parameters. P is one of the least reliably known band parameters, but because the tunnel rates do not vary strongly with P we can still expect to reliably "simulate" devices.

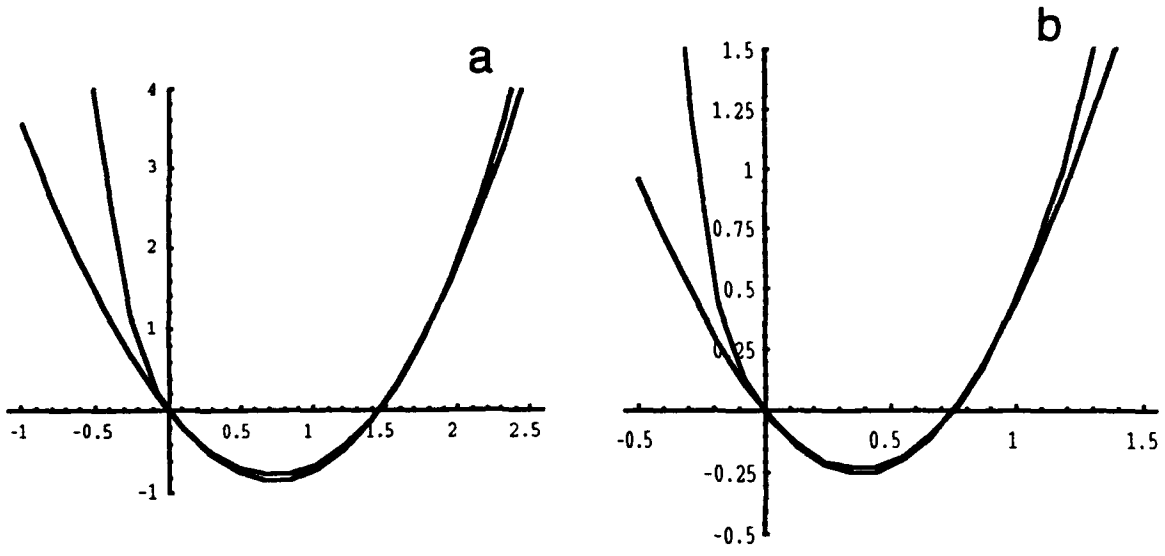


Figure 3

This compares the 2-band dispersion and the spherical Kane (8-band) dispersion using band structure parameters appropriate for InAlAs. This plot shows k^2 as a function of Energy by showing k^2, not just k, we can include both the allowed regions where $k^2 > 0$ and the "forbidden" region, $k^2 < 0$. Tunnelling will occur in the forbidden region. The tangent at $k^2 = 0$ is a line: , and similarly for the valence band. Because the 8-band model includes the split-off band, which "repels" the light hole band, there is a large non-parabolicity, and this is evident in the divergence of the two curves for negative E. For positive E, including the forbidden gap, the agreement is very good at the middle of the gap, the actual values are:

k^2= -0.762, for the 8 band case, and

k^2= -0.8419, for the 2 band case.

Figure b shows the similar structure for InGaAs.

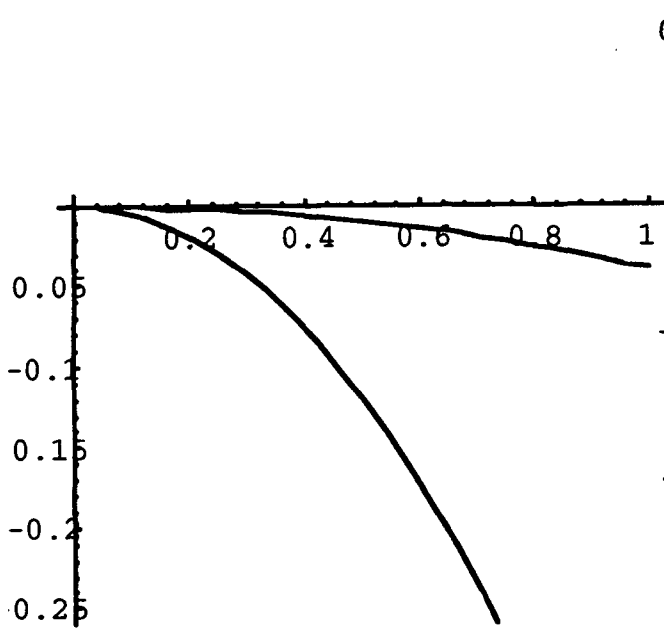


Figure 4a

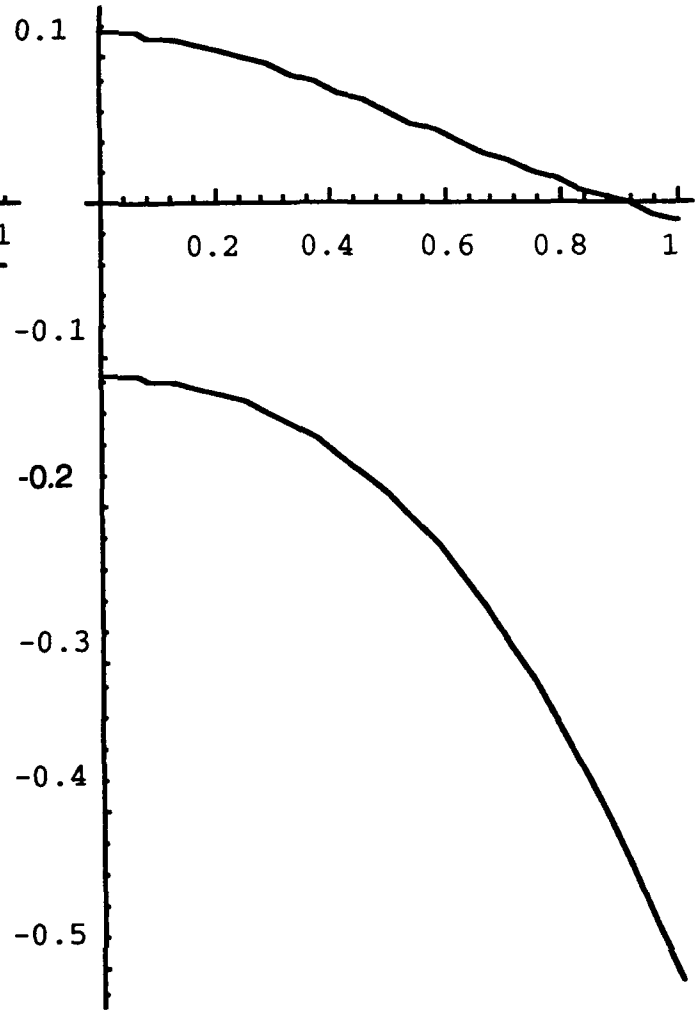


Figure 4b

This shows the dispersion in the 111 direction of light and heavy holes in GaAs. Figure 4a illustrates the case with no stress, and 4b shows the case in which a large stress in the 100 direction has split the bands by a total of roughly 200 meV. While this may be unrealistically large splitting, the qualitative effect remains the same for smaller splittings. This splitting would be typical of a distortion of several percent. This calculation used the Luttinger Hamiltonian, with stress terms, as discussed in the text.

Figure 4

The Full Kane Hamiltonian

Earlier, we wrote the spherically symmetric Kane Hamiltonian, and alluded to the fact that the full Hamiltonian contains band warping terms too. Before presenting a more complete Hamiltonian, we will motivate it's form.

Group Theory, and the Structure of Hamiltonians

The importance of symmetry to the theory of band structures dates was given explicit form in the paper of Bouckaert, Smoluchowski, and Wigner[14]. For our purposes, we need to construct the Hamiltonian from objects which are invariant under the cubic group. As the cubic group is a subgroup of the rotation group, any rotationally invariant Hamiltonian will be invariant. There is a special rank-4 tensor, which I denote T_4, which is also invariant under the cubic group:

$\mathbf{T}_4^{abcd} = \mathbf{x}^a\mathbf{x}^b\mathbf{x}^c\mathbf{x}^d + \mathbf{y}^a\mathbf{y}^b\mathbf{y}^c\mathbf{y}^d = \mathbf{z}^a\mathbf{z}^b\mathbf{z}^c\mathbf{z}^d$, where $\mathbf{x}^a$ is a unit vector in the x direction, etc. It is sometimes preferred to use the trace-free part of T_4, which is transformed as part of the irreducible representation of the rotation group, but is slightly more complex.

We can now add another term to the Hamiltonian equation 1, which acts upon the valence band "vector" components $\vec{\Psi}$. We can write

$$\left\langle \Psi^a \middle| \mathbf{H}_w \middle| \Psi^b \right\rangle = \gamma \mathbf{T}_4^{abcd} \frac{\hbar}{\mathrm{i}} \nabla^c \frac{\hbar}{\mathrm{i}} \nabla^d \qquad [3]$$

to indicate the action of the band-warping term in the Hamiltonian, where γ is a constant which defines the strength of the interaction.

When the electron spin operator $\mathbf{s}^a = \sigma^a/2$ is used, we can create several spinorbitwarping terms, but these are generally believed to be too small to be physically observable, and are rarely included in a Hamiltonian, so equation 3 is generally used as the sole anisotropic term.

The Luttinger Hamiltonian is the most general Hamiltonian allowed by group theory for the case of a set of four bands transforming as a spin 3/2 representation. This is a limiting case of the Kane 8-band Hamiltonian, with band warping, in the case where the momenta are all very small, so that the split-off band and the conduction band can be considered to be very far from the light and heavy holes. That limit is very appropriate for cyclotron resonance, where band parameters are often measured, and it is not far from the case of thermal carriers at room temperature, where the density of states must be determined. This is the most frequent starting point for the study of multiband systems, and virtually all tabulated band parameters are given as Luttinger parameters, $\gamma_1, \gamma_2 \gamma_3$. Using $ik = \nabla$. The Luttinger parameters are implicitly defined by the Luttinger Hamiltonian[15].

$$\mathbf{H}_L = \frac{\hbar^2}{m_e}\left[\left(\gamma_1 + \frac{5}{2}\gamma_2\right)\frac{k^2}{2} - \gamma_3(k\cdot\mathbf{J})^2 + (\gamma_3 - \gamma_2)\mathbf{T}^{abcd}\mathbf{J}^a\mathbf{J}^b k^c k^d\right], \qquad [4]$$

where $\mathbf{J}^a$ is the spin 3/2 angular momentum operator. $\mathbf{J}^a$ form a set of 3 four-by-four matrices. Unfortunately, some of the "simplicity" gained in the "elimination" of the split-off band is paid for in the form of these J matrices.

In the presence of stress, additional terms arise:

$$\mathbf{H} = \mathbf{H}_L + D_1\varepsilon^{aa} + D_2\mathbf{J}^a\mathbf{J}^b\varepsilon^{ab} + D_3\mathbf{J}^a\mathbf{J}^b\varepsilon^{cd}\mathbf{T}_4^{abcd}$$

$\mathbf{H}_L$ = Luttinger Hamiltonian [5]

The D's are deformation potentials. D_1 describes a term which shifts the overall band energy in response to isotropic pressure. This does not lead to large changes in the tunnel rates, and so will be ignored. D_2 is isotropic, in the sense that the relationship between band splitting and direction of stress is not dependent upon orientation. D_3 is similar to D_2, but the effect depends (through T_4) upon the orientation of the stress relative to the crystal axis. D_2 is expected to be larger than D_3, and since the actual parameters are not always well-known, we have assumed in the numerical work described later in this report.that the entire stress effect is due to D_2.

Effect of Stress

Stress has a significant effect upon the band structure of the zincblend semiconductors. At a qualitative level, the degeneracy of the valence bands is lifted. Figure 4 illustrates this, showing calculated results for GaAs. Notice that the upper band becomes very nonparabolic—this too is a general feature. Near the band maximum, the effective density-of-states mass is only 0.17, in contrast to 0.5 for the heavy holes. The stress terms enter into the Hamiltonian as described earlier in the theory section, and an 8-band simulator could incorporate these effects. We carried out numerical experiments in which the band offsets were varied, and found a very small effect. The main effect of stress is expected to be upon the valence-band density of states.

Our Tensor package, discussed in a following section will allow one to easily calculate the energies of the bands in an arbitrary stressed semiconductor, and this has been used to produce the graphics shown in Figure 16.

Simulations

We have developed a computer program to calculate the direct interband tunnel current, as illustrated in Figure 1. The direct thermionic current can (fairly accurately) be analytically estimated, and it is generally very small in the "valley" of low conduction. At the level of sophistication afforded by the present project, we have not attempted to calculate "indirect" currents, that is, currents that result from scattering into a bound state, and later from a tunnelling process through the barrier. We remind the reader that, as pointed out in Figure 1c, this indirect current can be the dominant current in a device.

Simulation Overview

It is usual to think of a device simulator as beginning with a set of equations (usually the drift diffusion equations and Poisson's equation), and then proceeding with a numerical solution. This works well in cases where there is agreement upon the device physics, but has caused frustration among device physicists who find that they are unable to embed new concepts or new physics into the context of a realistic device structure. I advocate a view of the device simulator as an environment which will set up a device band structure upon which the device physicist has the option of either solving "standard" equations, or else of embedding his own device physics. Because this is too ambitious a goal for the present purposes, we have implemented a simulator with a fixed set of physics, but have attempted to write code that will be applicable to a future, more general-purpose, device simulation environment. I would note, however, that this program has the capability to solver a simple Schroedinger's equation, using as the potential, the conduction band of a simulated device. This can be regarded as an example of utilizing the main program to provide an environment in the form of a realistic band structure, for the purpose of studying quantum effects upon charge distribution, and of testing a novel algorithm for the estimation of the Hartree (Quantum) charge.

Simulations can be broken into several parts: a Schroedinger's equation integrator, a Poisson solver, driver routines to produce J/V curves, etc:

In this proposal, we have been primarily concerned with the solution of Schroedinger's equation for the case of two coupled bands. This may be said to be the core of the project.

Before the Schroedinger's equation can be solved, the band diagram is needed. Schroedinger's equation can be solved for the resulting potential wells. In special cases, however, it is desirable to solve the Schroedinger's equation for simple wells. The computed results can then be compared with special cases for which exact analytic results are known. Generally, however, the actual band diagram is needed to predict the performance of a real device.

It is a conceptual complication of device physics that one tends to think of a process such as a particle tunnelling through a barrier in isolation, and yet actual devices are composed of many particles, and measurements observe averages. A device physicist with a good understanding of the underlying physics will often grasp a concept such as resonant tunnelling easily, but it is often far from clear how this process will affect a statistical average. Driver routines must direct the solution and presentation of a single quantum well problem (a single energy, for instance) when this is needed to help the user visualize the underlying physics, and then the driver must direct the solution of an ensemble of tunnelling problems and integrate the result to compare the theory and experiment.

Simulation Environment

Zytron's long-term interest is to develop software for a variety of platforms. Early versions of the user interface ran on both IBM-PC-compatible computers and Apple Macintosh computers. However, because we did not have the resources to maintain working graphic user interfaces (GUI's) on both machines, later development concentrated upon the Apple Macintosh, so that we would have at least one working GUI. The availability of a working GUI for numerical applications has proven very useful, and may lead to a standard form which can be transported to PC and UNIX-based computers.

At this time, the computer most used at ZYTRON for running simulations is a Macintosh IIcx with a Radius Rocket accelerator. This machine uses the 68040 RISC microprocessor, and is about as fast as any Macintosh system available at this time. One reason for our interest in transporting our codes to UNIX, of course, is that there is no upgrade path available to make computations upon a faster Macintosh computer.

Our present simulator would be expected to run on almost any Macintosh-compatible computer, although if the machine lacked a floating point processor, the program would need to be re-compiled to obtain a non FPU version. A non FPU version would frankly, be so slow as to not be very useful.

Mathematical Models

In this proposal, we have been primarily concerned with the solution of the Schroedinger's equation for the case of two coupled bands, which can be written in the matrix form discussed in the previous section and written as equation 2.

This may be said to be the core of the project. To calculate transmission through the barrier, by tunneling, one needs the transmission coefficients, which can best be defined by referring to Figure 5, which indicates the proper normalization, with a graphic definition of the various waves. A complete linearly independent set of solutions to the Schroedinger's equation can be used to create a solution as defined in the figure, but in practice some solutions are close to being linearly dependent and lead to large numerical errors.

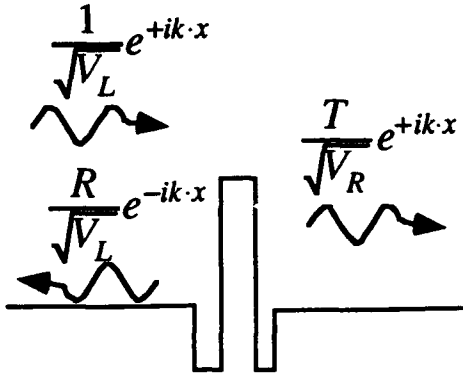


Figure 5

This illustrates the normalization for the calculation of a transmission coefficient. The wavefunction should be normalized for unit flux, not unit amplitude. In multiband models the velocity is

$$V = \hbar \frac{dE}{dk}, \text{ and not just } k.$$

Each incoming wave transmits independently of the others, so the total current is

$$J = \frac{gq}{2\pi\hbar} \int \frac{d^2k_p}{(2\pi)^2} \int dE |T|^2 (f_L - f_R) \quad [6]$$

g = degeneracy = 2 and f_L is a Fermi function

$$f_L = \frac{1}{1 - e^{\beta(\mu_c - E)}}$$ Fermi Factor and similarly for f_R.

As has already been pointed out, the band diagram is needed before the Schroedinger's equations can be solved.

Numerical Methods

Almost all of the core numerical routines used in this project trace back to routines in the book "Numerical Recipes in C"[16]. The most significant addition, made, was a set of routines for banded matrices, which are however based upon routines in Numerical Recipes for the upper-lower decomposition, and the "backsubstitution" for general matrices.

Numerical Difficulties

The Schroedinger's equation takes the form of a set of two coupled (ordinary) differential equations. The solution of differential equations is often regarded as straightforward, but for our purposes there are two difficulties which deserve special attention. The first concerns the fact that in some cases the actual equation is not well behaved. Figure 6 illustrates a pathological band diagram. In a real device a particle "propagating" at the forbidden energy E_t will consist of an exponentially decreasing wave. In the case of figure 6, however, the wave can propagate unattenuated with momenta at point Q. The result is that the transmission coefficients always take values close to 1, and the computed tunnel current is very large. While this is a pathological example, such situations can and do arise when one attempts to find a "best" fit to the band structure at the Γ point. The X point, which is the boundary of the Brillouin zone in the Δ direction is near 11/nm, and so the pathology shown by this example is far out in the band diagram. In a real global band structure higher-order effects keep the gap open. The discretization can mitigate this problem by truncating the available configuration space to include only low k values, but the results will depend upon the details of the discretization. Essentially, this amounts to the use of a tight binding Hamiltonian. In our case, we forbade the use of band parameters which close the gap. It is easy to show that the gap will be open if

$$A_c A_v > 0.$$

Before leaving this subject, I want to emphasize that this pathology is not a feature of the solution method, but instead is a pathology of the band equations themselves. This pathology can arise in the 8-band model as well as in the 2-band model used in the present research.

The second difficulty has received attention in the literature McGill, et. al [7], but is easy to fix. A two-band Hamiltonian will have two

solutions, one of which will generally grow at a large exponential rate. If the equation is integrated using a differential equation algorithm, such as Runge-Kutta, then the fast growing solution will overwhelm the other. Unfortunately, the physically-relevant solution, is the slowly growing solution. It can happen that the precision of the machine is inadequate to uncover the relevant slowly-growing (or oscillating) solution. When phrased in terms of linear algebra, this problem can be viewed as the result of using an unstable ordering to solve the equations (the importance of pivoting in the solution of linear equations is discussed in length in many textbooks, such as Numerical Recipe's). Our solution to this problem is to use an ordering in which the diagonal matrix elements are used as the pivots. The Schroedinger's Equation is diagonally dominant for "small" energies, where "small" is relative to the maximum, and so this works well. McGill *et .al.* use a tight binding approach, but I believe that it is in actuality very similar to our method. This problem and solution are well known in the literature of device simulation.

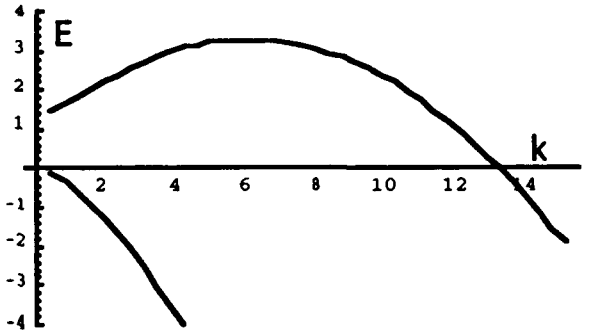


Figure 6

This shows the band diagram for AlInAs, in which the band parameter a2 has been set to 0.05, while the other parameters retain standard values. Normally a2 and e2 are both negative. This change has a very small effect upon the band structure near the Γ point. For very large k's, the conduction band has a maximum and bends down. The result is that there is no gap.

Representation of Differential Equations

Three different differential equations arise in this research, but they can all be solved using essentially the same method. We use a finite element method, which can be viewed as using a set of "tent functions" as basis elements (see Figure 7). The Hamiltonian H (or Laplacian, L or inner product K) are all computed by performing integrals:

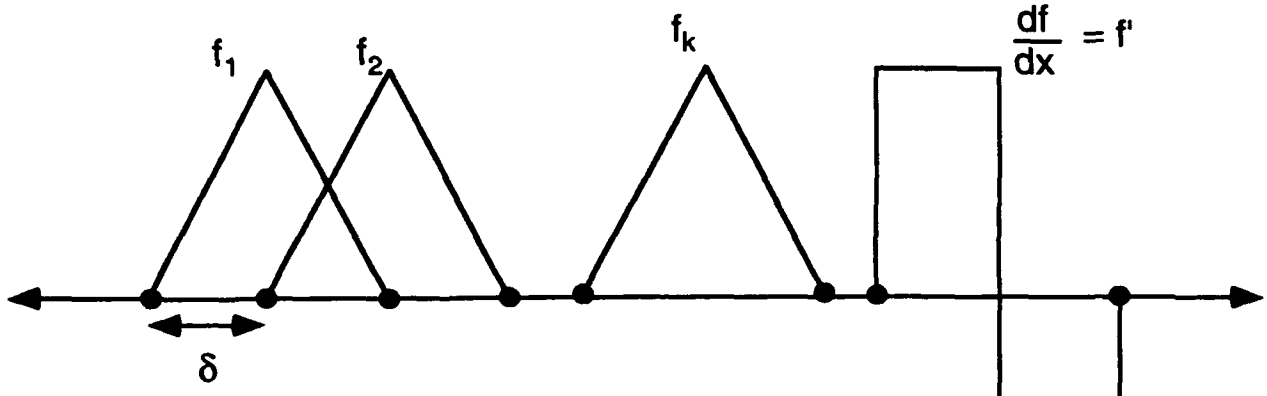
$$K_{ij} = \langle h_i | h_j \rangle = \int h_i(x) h_j(x)\, dx$$

$$L_{ij} = \langle \nabla h_i | \nabla h_j \rangle = \int \left[\frac{dh_i(x)}{dx}\right]\left[\frac{dh_j}{dx}\right] dx$$

H requires additional terms but is similar. Figure 7 lists some of the needed integrals. In our basis set, these are all banded matrices. Computation of the A-band matrix equation solver can then be used to find the potential, or Green's function. The form shown here for L, in which a first-order derivative is applied to each element, instead of a second derivative being applied to the right element, is superior in two ways: first, it is generally easier to compute, and second, it is manifestly self-adjoint, while the double-derivative form is only self-adjoint if the boundary conditions are properly handled.

One may note that this form of the differential equations has no obvious boundary conditions. If we simply proceed, we find that in fact the equations amount to Neuman boundary conditions, in which the derivative of the function is zero. Dirichlet boundary conditions, in which the value is defined, can be obtained too. To specify Dirichlet boundary conditions, one must modify the equations. First, the values at the ends are no longer "unknowns," and so the number of equations must be reduced by removing the equations for the end points. The values of the end points will then give rise to an inhomogeneous term for the nearest neighbors to the ends.

In our simulations, the Schroedinger's equations were solved for both types of boundary conditions, with the user being able to select which solution is desired. The Poisson's equation was always solved for Dirichlet boundary conditions because L has a zero eigenvector (constant potential not zero), which makes L^{-1} undefined.



$$\int (f_1)^2 = \frac{2\delta}{3}$$

$$\int f_1 f_2 = \frac{\delta}{6}$$

$$\int f_i f_k = 0$$

$$\int \left(f_1'\right)^2 = 2\delta$$

$$\int f_1' f_2' = -\delta$$

$$\int f_1' f_2 = \frac{-\delta}{2}$$

The basis functions used in this work are piecewise linear, and can be considered to be a linear superposition of the "tent functions" illustrated here. Matrix elements involving piecewise linear potentials or derivatives are easy to compute using the accompanying formulae. These basis functions turn out to be non-orthogonal, so that the inner product must be computed, but this turns out not to be a large complication. The matrix elements can be derived from simple formulae, as shown. The actual program used a generalization, to allow for nonuniform grid spacing.

Figure 7

An advantage of the finite element method over finite differences, or over tight binding methods, is that it is easy to use an irregular grid. In our case, the device is first defined as a set of regions, one for each material, with discontinuities (in the band edges) allowed at the boundaries between regions. Our grid routine places grid points at the boundaries, and then adds additional points until a total number of grid points is reached.

A complication of the two-band model is that the wavefunction is defined by two amplitudes at each point, not one. In the case of an n-band model, one needs n amplitudes at each point, one for each band. Our present program represents the wavefunction upon an N point grid as a single 2 N long vector, with v[1] and v[2] representing the conduction and valence band components, respectively for the first gridpoint. v[3] and v[4] represent the same bands at the second grid point, etc. In order to make the text more clear, however, I will write as though there was a separate length-2 vector at each grid point so that the conduction and valence band amplitudes at the first point will become c[1] and v[1].

Poisson's Equation Solver

Our Poisson equation solver assumes that the carrier densities are in equilibrium, but with a separate chemical Potential for electrons μ_c and for holes μ_v so that the carrier concentrations can be written as:

$$n_c = \left[\frac{g}{(2\pi)^2}\left[\frac{2m_c}{\hbar^2}\right]^{3/2}\beta^{-3/2}\right]F_{1/2}\left(\beta(\mu_c - E_c + \phi)\right)$$

$$n_v = \left[\frac{g}{(2\pi)^2}\left[\frac{2m_v}{\hbar^2}\right]^{3/2}\beta^{-3/2}\right]F_{1/2}\left(-\beta(\mu_v - E_v + \phi)\right)$$

[8]

while ϕ obeys Poisson's equation

$$-\nabla\varepsilon\,\nabla\phi = \rho = q\left(n_d + n_v - n_c\right) \qquad [9]$$

Real equalibration would require $\mu_c = \mu_v$, but equation 3 can be solved for any μ_c, μ_v pair, and gives a physically reasonable solution when $\mu_c - \mu_v < E_g$ the band gap. Straightforward algebra shows that if n_c and n_v obey equation 3 with $\phi=\phi_g$, then one iteration of Newton's method applied to finding a better ϕ is equivalent to solving the equation:

$$-\nabla\varepsilon\,\nabla\phi =$$

$$q\left[n_d + n_v - n_c + \frac{dn_v}{d\phi}\left(\phi - \phi_g\right) - \frac{dn_c}{d\phi}\left(\phi - \phi_g\right)\right]$$

[10]

$$\frac{dn}{d\phi} \quad \text{Taken at} \quad \phi = \phi_g$$

where ϕ_g is our initial guess for ϕ, and to calculate the derivative we will use the fact that

$$\frac{dF_n(z)}{dz} = nF_{(n-1)}(z)$$

Newton's method converges rapidly if the initial guess is close to the final solution, but can diverge if it is not close enough. We start with the carrier concentrations equal to the doping, except where the doping is very small, in which case we put $\phi = (E_C + E_V)/2$. For our quasi-equilibrium case, it is possible to guarantee convergence. We simply compute the free energy of the device for the new ϕ. If the free energy does not decrease, then we know that our new ϕ is not better than the old. If the free energy increases, then we can try a new $\phi_{new} = (\phi + \phi_g)/2$. This trick to guarantee convergence worked wonderfully when I applied it to an FET simulator several years ago. In this case it was rarely needed, as Newtons method almost always converged unaided. One can note that the fermion free energy density is given by

$$G = -\frac{g}{(2\pi)^2}\left[\frac{2m}{\hbar^2}\right]^{\frac{3}{2}}\beta^{\frac{-3}{2}}\frac{2}{3}F_{\frac{3}{2}}(\beta v) + nE_B$$

$$+\frac{1}{2}\phi\rho + \text{Boundry Terms}$$

11a,

where $\upsilon = \mu - E_B + \phi$.

A separate contribution must be added for each band, and to avoid counting the electrostatic energy twice, E_b should not include the electric potential. One can also write the total electrostatic free energy as

$$F = -\frac{1}{2}\varepsilon(\nabla\phi)^2 + \phi\rho = \frac{1}{2}\phi\rho + \quad \text{Boundry Terms} \qquad 11b,$$

which defines the boundary terms in eq. 11a. This equation can be used to define the boundary terms. Unfortunately, equation 11b is not positive definite, so that the solution to Poisson's equation itself must be used to compute the free energy, not the approximate solution obtained from iterating Newtons method. We always used Dirichlet boundary conditions for finding ϕ, but one could use Neuman boundary conditions for ϕ if the Newtons method is used, and the free energy is not checked, as equation 10 is well defined for Neuman boundary conditions, unlike Poisson's equation.

Last, for the solution of Poisson's equation using Dirichlet boundary conditions, one must specify ϕ on the boundaries,. We set ϕ such that the carrier concentrations would equal the doping, so that everything behaves smoothly.

Once the equations are in the form of linear equations, it is easy to solve them numerically. Newtons method is used, and iteration is stopped when the potential ϕ_g is very close to the potential ϕ obtained by solving Poisson's equation using the charge distributions defined by ϕ_g.

Calculation of Transmission Coefficients

As described above, the finite element gives us a representation of the Hamiltonian, H, and the inner product, which we will call K. If our basis were orthonormal, K would be the identity matrix, but that is not the case here, where K is a banded matrix. As discussed above, it is straightforward to solve the linear equation for G:

$$(H - Ek)G_k = c_k \quad \text{or} \quad \nu_k \qquad [12a]$$

and we refer to G as a Green's function, because we will generally put

$$c_k[i] = 0 \qquad i \neq k$$

$$c_k[k] = 1 \quad \text{Similarly for } \nu_k \qquad [12b]$$

The simulation package will allow the user to set k to any integer and view the Green's function, but for the purpose of computing the transmission coefficients, we will only use k = 1 or k = N, the length of the grid. Setting k= 1 and k=N, the Green's function solutions give us a set of 4 (at each point each band gives a solution) wavefunctions which obey equation 12, inside the device. One can then consider the constraints of no ingoing waves, except for an ingoing wave of unit amplitude. This solution method incorporates a boundary condition on both the left and the right, and as a rule only exponentially decreasing functions are obtained. As a result, the matrix operations needed to calculate the transmission coefficient T, and the reflection coefficient R, if desired, are well conditioned.

Eigensystems—Calculation of Wavefunctions and Energies

In order to calculate the transmission coefficients, one must create the Hamiltonian and the inner product (our basis set was not normalized). The Numerical Recipes routines for the solution of general (symmetric) eigensystems was used to calculate energies and wavefunctions. This was expedient to do, as I had heavily used the routines in the past, and had written the Graham Schmidt orthonormalization routine needed to create an orthonormal basis set. This is not very efficient, for our instance, as it does not take advantage of the banded structure of the Hamiltonian, or the fact that often only the lowest few eigenvectors are needed. It was not expected that the eigenanalysis would be used very much, so optimization could not be justified.

Calculation of Fermi Integrals

The charge densities and free energies used in this work have required the calculation of Fermi integrals:

$$F_n(z) = \int_0^\infty \frac{x^n dx}{1+e^{(x-z)}}$$

and of the inverse function. Note that this differs by a factor of $\Gamma(n+1)$ from a commonly-used convention for these integrals. Our simulator includes a package to compute such integrals. If the argument, z, is very large, then an asymptotic formula is used; otherwise, the package simply carries out a numerical integration. Numerical integration is too inefficient for those orders which are heavily used, so that for certain orders a table is tabulated. Those orders which are accessed often (hundreds of times by the Poisson solver for orders 1/2 and 3/2, for example) do not result in constant reintegration. Our routine will work for negative (non-integer) orders, as it treats the region near x=0 by making an analytic approximation that allows us to access order -1/2. Order 0 is treated as a special case, as

$$F_0(z) = \log\left(1+e^z\right)$$

Inverses were not needed often, and we used a Numerical Recipes routine based upon bracketing and bisection. This, of course, could be dramatically speeded up if needed.

As a final digression, we add our own contribution to the long list of approximations to the Fermi integrals for orders 1/2 and 3/2. These have the property that they are easily invertible and exactly reproduce the asymptotic behavior of the exact integral. The form for order 1/2 is accurate to an astounding 1% for <u>all</u> values of z, while the form for order 3/2 is off by 7% at the worst. The forms are:

$$F_{1/2}(z) = C\left[\log\left(1+Ae^{2z/3}+Be^{4z/3}\right)\right]^{3/2}$$

$$C = \frac{\sqrt{3}}{4}$$

$$A = \left(\frac{\sqrt{\pi}}{2C}\right)^{2/3}$$

$$B = 1.22$$

$$F_{3/2}(z) = C\left[\log\left(1+Ae^{2z/5}+Be^{4z/5}\right)\right]^{5/2}$$

$$A = 1.2934$$

$$B = 1.1$$

$$C = 0.69877$$

Inverting these formulae requires solving a quadratic equation to obtain the exponential term.

Note on units

It is surprising how often the units for physical quantities can cause difficulties to calculation, even for "experts." Many papers use units such as "atomic Units," which eliminate "constants" but are often difficult to translate into "normal" units. Worse, "bastardized" units, which combine some MKS, some CGS, and occasionally some non-metric as well, are common.

The convention which I have found to be best is relatively simple. It works well to base units upon MKS, or scaled MKS. This is in contrast to atomic units, which make some fundamental constants equal to 1. The utility of such units is especially doubtful in solid-state physics, where so many "constants," such as ε, are dependent upon the material anyway. In this work, the unit of length is the nm, for example. One advantage is that all the fundamental constants are tabulated in MKS and can be used directly. There are however special exceptions, the primary one being the use of electron volts for the microscopic unit of energy. "Microscopic" means that it refers to a single particle. In device physics, other

energies sometimes occur, such as power dissipation, or capacitive energy. These macroscopic energies are best kept in pure MKS units.

As a special help in keeping units straight, I generally introduce a set of variables called UNIT_L. UNIT_E, etc., which hold the MKS values of all units in use (length and energy in these cases). With this convention, one can scale any item correctly and even change the units in use later.

Note on Notation

We have generally attempted to follow common notation in this report, but in some cases there is no universally-agreed-upon notation, and in at least two instances our notation is known to be at variance with frequently used notation:

Our notation for Fermi integrals is

$$F_n(z) = \int_0^\infty \frac{x^n dx}{1 + e^{(x-z)}}$$

while others often use

$$F_n(z) = \frac{1}{\Gamma(n+1)} \int_0^\infty \frac{x^n dx}{1 + e^{(x-z)}}$$

I frankly fail to understand the motivation for the Γ function; it appears to only make the later algebra more complex and error prone.

Some writers have used T to denote the square of (our) transmission coefficient. I would point out that the subject of tunneling is the aberration; in other disciplines, ranging from high-energy physics, to microwave engineering, one generally speaks of a scattering amplitude, which is a complex number, and the flux of scattered energy varies as the square of the amplitude. Our notation is thus in agreement with the greater body of physics and engineering literature.

User Interface

In twenty years of programming to solve numerical problems, I have observed that input/output issues lead to an inordinate amount of difficulty. Most programmers tend to do what comes easily, which makes their programs very hard to use. This contract has provided an opportunity to experiment with a graphical user interface for numerical work. We have tried to make the interface easy to program, so that it could be kept "up to date" as the numerical goals changed. In this, our interface differs from the traditional graphical user interface (GUI), which typically requires great time and expense for each problem. In a "typical" GUI application, such as MacDraw, the interface in fact comprises most of the program, and is certainly the most difficult part to implement.

The User's View

Other than launching the application, all user interaction and file I/O is carried out through the user interface. As much as possible, I have attempted to conform to the Macintosh user interface guidelines, and published in "Inside Macintosh"[17] and elsewhere. These are, in fact representative of good GUI principles, and so this is not a great loss of generality. User interaction takes three forms: selecting pull-down menus, viewing plots, and editing variables.

At present, the plots are generated as part of commands to perform a computation; they take very little time to produce, and can be immediately deleted by a user who does not want them. Each plot appears in its own window, which can be dragged, resized, or deleted, independently of any other plots. The ability to drag plots makes it easy to compare different ones, such as the two-band diagrams shown in Fig 8. Fig 8 is a "screen dump," showing a set of six representative plots The plots in Figure 8 are as follows: The upper-left plot shows the band diagram and carrier concentrations for a double-well RITD with two 4nm wells separated by a 4nm barrier. The other plots all concern a similar device with a 6nm barrier. The remaining plots are, in clockwise order, from the upper right

The transmission coefficient verses energy of the incoming electron (broadest peak). The two other curves are the Fermi function difference (rightmost peak) and the product $|T|^2 (F_l - F_R)$ --- the sharp peak in the middle.

Quantum wavefunctions. The conduction band can be used as a potential to solve a simple Schroedinger's equation. This plot shows the band, and the first six energy eigenfunctions, with the boundary condition of zero derivative at the ends.

Charge densities. This plot compares three charge densities. One is the Fermi-Thomas charge density, the others are the Hartree charge density, and the last is an approximation to the Hartree density, which is much easier to compute. The approximation is very close to the Hartree density, but is slightly larger. The details of the approximation are discussed in the text.

Green's Functions. This shows a Green's Function, a solution to the Schroedinger's equation, with an inhomogeneous boundary condition. The two curves show the two components of a 2-band wave function. The wave function does penetrate through to the right, but is difficult to see, as the tunnel coefficient is only of the order 0.01 . Also shown is the band diagram.

Band Diagram. This is the band diagram, and carrier concentrations for the double-well RITD, with a 6nm barrier. Also shown are the carrier concentrations. Note that it can be compared with the band diagram of the 4nm barrier device, which appears directly above it.

resulting from device simulations. Plots are intended to give rapid feedback to the user of what the simulator is doing, and are not intended to be "publication-quality". Publication-quality graphics packages tend to be much slower and to use more memory. The stored output of our simulator can be used as input to software that produces better graphics, if desired. Many of the figures in this report were produced by reading data into *Mathematica* and then creating plots.

The Editor

The most unique feature of our user interface is a special editor for numerical variables used by the programs. Variables are listed in a window, which scrolls if needed. This feature of the user interface can be described as a great success, in that it improves the productivity of the user substantially. Figure 9 shows how the computer screen appears, when using the editor.

Pull-Down Menus

The pull-down menus allow the user to specify any one of a set of commands:

The **File Menu** allows the user to save variables in one of two AASCI-based formats, and to read the variables back in again. The standard format is designed to allow the same variables to be read back in by the simulation program. As an alternative, the variables can be saved as a *Mathematica* file. In this case, the file can be read by the program *Mathematica* , and the variables will all be defined with the same names used in the program, and the same values as when they were first saved. In this case, the powerful graphical and transformation capabilities of *Mathematica* can be used to analyze the data. At this time, variables can be integers, real numbers, or vectors.

The **Quantum Menu** is the main menu for directing the solution of equations. A device, or quantum well, must already be defined in order for the choices on this menu to be available. In this case, one can:

Solve the Poisson's equation, self-consistently, for a device. The charge density will be given by a Fermi-Thomas approximation with finite temperature.

Compute Quantum wavefunctions for a quantum well. At this time, a simplified system is solved using only the conduction band, and using a constant mass. Nonetheless, the solutions are useful to help visualize the device physics. The quantum well can be the result of solving the Poisson equation, and so it can be realistic. The wavefunctions can be

used to commute the "quantum charge," which is the "exact" charge in the Hartree model, assuming band-filling obeys usual Fermi Statistics.

Quantum Charge is available only after the wavefunctions have been computed, and will calculate the charge density. This menu choice also calculates and displays an approximate charge density, to be described elsewhere.

Green's Function will calculate a Green's function for the Schroedinger's equation. The Green's functions are used to calculate the transmission coefficient, but are recomputed when needed (the computation is fast).

Transmission Curve will calculate the absolute value of the transmission coefficient, $|T|$ verses energy, for a single one-dimensional quantum well. Also computed, and displayed, are the difference in Fermi coefficients (F_l - F_R), which weight the energy by the occupancies on the left and right, and the product $|T|^2$ (F_l - F_R), which is the integrand that contributes to the net current.

The Transmission vs Kpar choice will integrate the current at a fixed bias voltage, and display the integrand as a function of the parallel momentum (Kpar). Thus each point on the integrand curve represents an integral over energy.

The J/V curve choice performs a double integral over energy and parallel momentum for a range of bias voltages, and displays an entire J/V curve.

The **Variable-Edit Menu** is used to start a new instance of the editor, which will be discussed later.

The **Junk** menu is a set of routines that are not closely related to device simulation (although one will plot Fermi Integrals). They were used as test programs for early versions of the human interface.

The **Device Menu** causes a device data structure to be defined. Roughly 10 devices and quantum-well structures are defined at present, representing all of those used in this work. A device-creation routine can reference variables that are editable in the v-editor, and so a single device routine is all that is needed to create any one of a large set of devices, such as any double-well Resonant interband tunnel diode.

Programming the Interface

When the design of this simulator was begun, it was hoped that it would be possible to design the interface in such a way as to make it very easy to program. While, the interface system is "easy" to program, relative to alternatives, full attainment of our goal was limited by the present state of the art of software tools. It now appears almost certain that standard c does not allow attainment of this goal, and the use of a pre-processor would appear to be indicated. A serious problem with c is that the same (or similar) information must be manually entered in as many as 5 places in the program. This causes an obvious waste of human effort, at best, and can be a source of serious programming errors as well. A summary of the entries needed is:

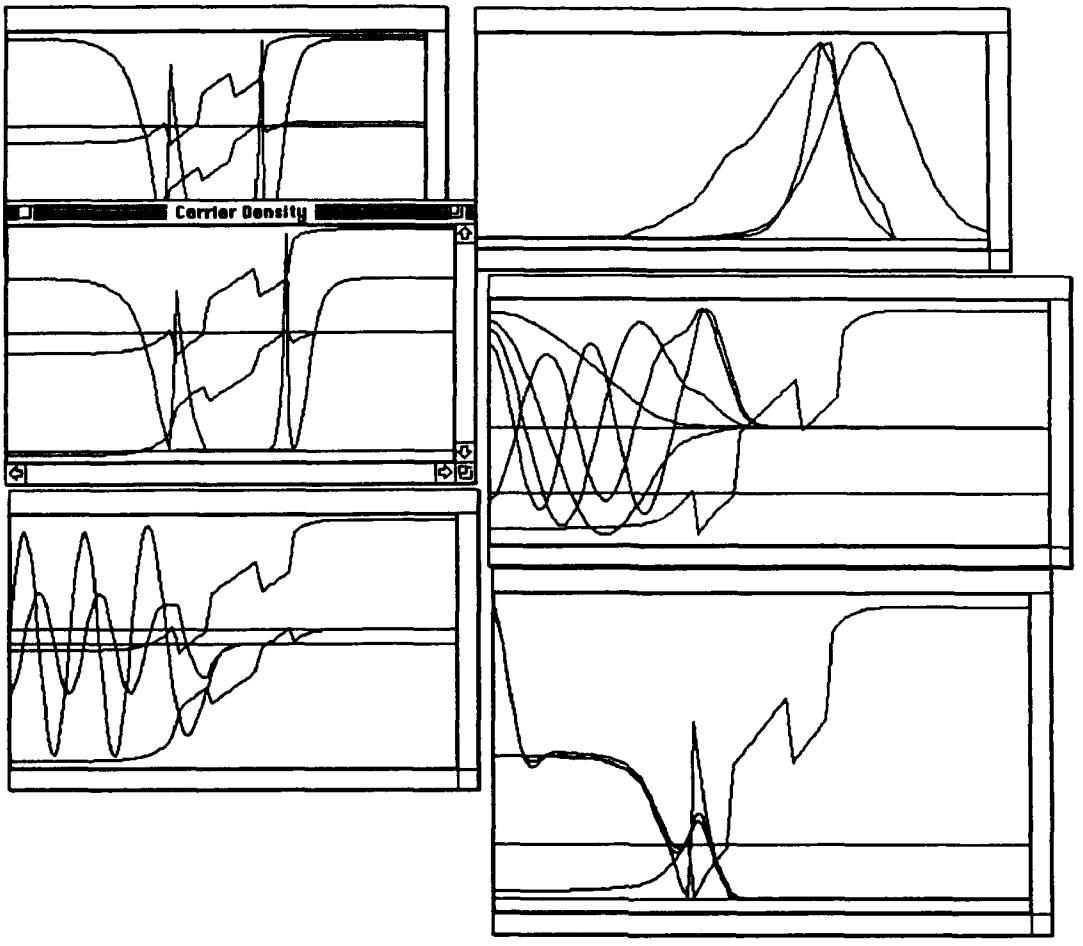
1. Variable declaration.

2. Create a data structure: The editor must be aware of the variable and its properties, including its default value and type. For this to be the case the information must be placed into a data structure, which is handled by calling a subroutine

3. Allocate memory: If the variable is a vector or matrix, storage must be allocated for it.

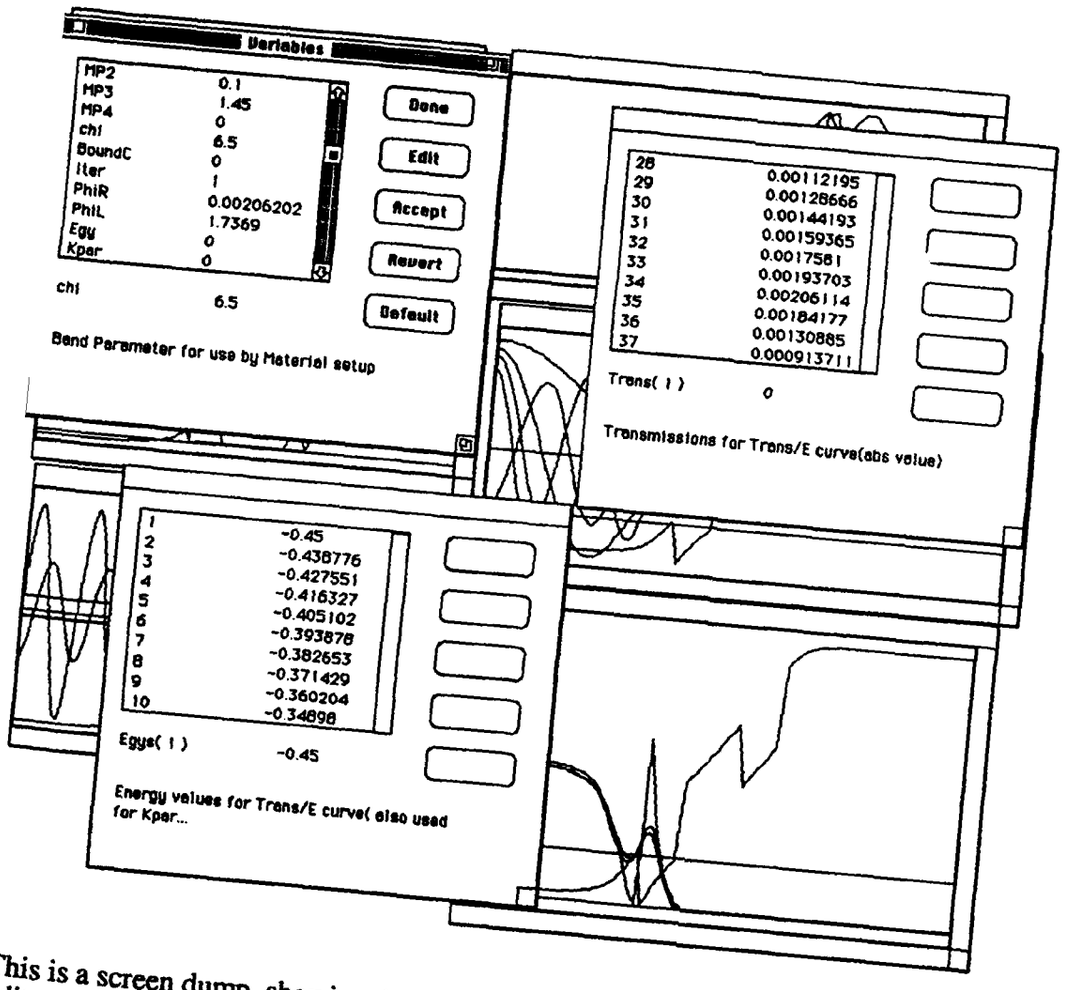
4. Free memory.

5. Update variable: At present, the editor knows (from the data structure set up in 2) the location in memory where the variable is stored, and automatically incorporates changes made by the user. There is thus no need for independent programming for each variable; however, this arrangement makes it too easy for the user to destroy the integrity of the data structures. The user can, for example change



This shows a "screen dump," made while the ZYTRON device simulator was running. It shows a copy of the actual computer screen, except that the original is in color. In transforming to a black-and-white form, the graphs become more difficult to read than they are on a color monitor, and some features, such as labels upon the "unselected" plots do not appear. By clicking and dragging with a mouse, the user can move, resize, or delete plots. In normal operation, plots tend to overlap to a greater degree than shown, but are redrawn when selected or uncovered. In this figure, only the upper-left plot is significantly obscured.
The upper-left plot shows the band diagram and carrier concentrations for a double-well RITD with two 4nm wells separated by a 4nm barrier. The other plots all concern a similar device with a 6nm barrier. The remaining plots are described in the main text

Figure 8



This is a screen dump, showing the appearance of the variable editor. The upper-left window is the main editor window, showing the variables and their values. The scroll bar can be used to move through the entire list, while the buttons along the right send commands to the editor. When a variable is selected (using the mouse to point), the name value and a descriptive text appear, as is shown for the variable chi. One can change the value of the selected variable by typing in a new value. If a vector is selected and the edit button pushed, then another window appears, so that individual elements of the array can be examined or changed. Vector windows remain open until closed, and the screen dump shows two open.

In this figure, the editor windows are overlaid on top of a set of plots.

Figure 9

the value used to indicate the dimension of a vector, without the vector being reallocated to reflect the change.

By means of a preprocessor, the relevant information could be entered only once, and all other entries would be made in a consistent manner.

Setting up the menus also required not only that compatible information be entered in several parts of the program, but also that a utility program called ResEdit be used to create the entry in the menu bar. Programming the menu system was further complicated by the fact that not all menu choices are relevant all the time. One should not, for example, be able to solve for a band structure before a device structure has been set up. In the present implementation, a set of flags indicate the status of the data structures, but the nature of this problem was not appreciated until a version of the simulator was running and could be tried out. A more systematic approach should make setting up the enable/disable flags much more straightforward.

On the other hand, programming the plots proved to be easy, as there are only a few subroutine calls which need to be made—one to create the plot, one for each curve to plot, and one for special options such as forcing two curves to use a common coordinate system, or drawing a line at 0. Other plotting subroutine packages which I have used require large numbers of parameters to be specified, such as the location of tickmarks, labels, colors, etc. My present philosophy is that reasonable defaults can be made so that none of these parameters are needed, and one can allow the user to change the plot style if they wish.

Special Tricks

Several special tricks have been added to what were essentially Numerical Recipes conventions for numerical computing in the C language. This section presumes a familiarity with C, and with the material in Numerical Recipes in C, referred to as NR for the rest of this section.

Precision has tended to cause trouble in C, because the C convention wants to convert everything to a double prior to carrying out operations or subroutine calls. While this has been alleviated somewhat in the ANSI standard, there are still compilers for which operations on floats are slower than on doubles. In most of our code we use a type FLOAT, which is to be defined in math.h . Thus, the type can be changed easily and in a global fashion.

Extensions to the vector and matrix routines have been added. The Numerical Recipes Routines for vectors and matrices are extremely useful, and we have made several minor "extensions." Vectors and matrices are represented as pointers. One extension is to allocate a few extra bytes at the start of the object. The actual value of the pointer can still point to the start of the data, so that this change is compatible with all of the standard NR routines. These extra bytes can be used to store information about the object, such as the length. The main use, in fact, is that there are length functions which can be applied to (most) vectors, and matrices, and which will return the dimensions. At present, the system only really works on standard origin-1 objects, but there appear to be ways to improve this.

A banded matrix type has been added. In general, we have delayed detailed optimization of our code until future versions, but the simulator needed some band-matrix routines very badly, and there is a clever way to store banded matrices in C, which (to my knowledge) is not widely used. In brief, the trick is to modify the NR matrix routine, so that the allocated column is the width of the non-zero part of the column. The special trick is to offset the pointer, so that element $M_{i,j}$ is referred to as M[i][j]. Thus, a routine need not really take into account the banded nature of M, except, of course, that it must not refer to the non-allocated elements of M.

New or Original Work

The material in this section is varied, but is characterized by it's being substantially different from other approaches taken to solve similar problems. Of course, symbolic manipulators have been heavily used in physics, and our Tensor package is modeled after the γ matrix packages designed to help high energy physicists cope with the complexity of the Dirac equation. The methods under development for the estimation of the Hartree charge density is, to my knowledge highly unique, but similar methods may have been used at some point in solid state physics.

Spin 3/2 Tensor Package

The theory of the valence bands of zincblend semiconductors has been plagued by algebraic complexity since the mid 1950's, when the need arose to incorporate both spin orbit splitting, and band warping into the Hamiltonian which describes the band structure. On the one hand, the equations are complex and their manipulation by "manual" means is slow and error-prone, while the equations are also characterized by a finite set of operators, closed under (non-commutative) multiplication and addition. These facts suggested that computer manipulation would be an appropriate tool to advance the formalism used to describe the physics of the valence bands. With this in mind, we implemented a special tensor package in *Mathematica* which will manipulate the tensors and non-abelian operators arising in complex multi-band theories, such as the Luttinger Hamiltonian description of valence bands.

Operator Algebra

The core of the tensor package, is made up of operator reduction routines to manipulate the spin 1/2 and spin 3/2 angular momentum matrices. Our package is unusual in that it is "basis-free." Basis-free calculation has been used for many years by high-energy physicists to manipulate the Dirac γ matrices. Two simple formulas suffice to reduce <u>any</u> expression involving the **J** matrices to a standard form. Our standard form utilizes the fully symmetric products of **J** matrices, as the fundamental objects, so that we have,

the 3 **J**'s themselves $\mathbf{J}^a$

5 symmetric trace free products

$$\mathbf{J}^a\mathbf{J}^b - \frac{1}{3}\delta^{ab}\mathbf{J}^2$$

and

7 products of 3 **J**'s .

Together with the identity matrix, this gives us all 16 possible 4-by-4 matrices, in accordance with group theory's assurance that there can be no symmetric trace-free product of 4 or more spin 3/2 operators. The result is quite general—in the case of spin n/2, there can be no symmetric trace-free product of n+1 or more operators.

First we reduce all unsymmetrized products to a fully symmetric form using the commutator:

$$2\,\mathbf{J}^{[a}\mathbf{J}^{b]} = i\,\varepsilon^{abc}\mathbf{J}^c$$

where square brackets denote antisymmetrization. The remaining fully-symmetric products constitute our standard form, and of course all products higher than 3 (1 for spin 1/2) are set to 0.

The sum of diagonal elements of a matrix is independent of the basis, and is usually called the trace, but to avoid confusion with the trace mentioned above, contraction with a δ, we will call it the spur. In any event, the spur becomes trivial to compute, as the spur of a trace-free product of **J**'s cannot be non 0. This is a result of group theory, as the spur is a scalar, under rotations, while a trace-free product of **J**'s transforms as an irreducible (non trivial) representation of the rotation group.

Other Tensors

In order to manipulate expressions involving **J** operators, the tensor package must also manipulate the antisymmetric tensor $\varepsilon^{\mu\nu\rho}$, and the delta, δ^{ab}. Finally, the tensor package will

manipulate the tensor T_4, invariant under the cubic group. Additional invariant tensors also can arise, as T_4 's are contracted, but a total of two can be used to make a complete set.

Application of Tensor Package

As an example of the application of the tensor package, we can take the spur (trace) of the Luttinger Hamiltonian:

$$\text{spur}\left(\mathbf{H}_L\right) = \frac{\hbar^2}{m_e} 2\gamma_1 k^2$$

Taking the trace of the square, using the tensor package was easier than typesetting the resulting equation:

$$\text{spur}\left(\mathbf{H}_L{}^2\right) = \left(\frac{\hbar^2}{m_e}\right)^2 \left\{\left(\gamma_1{}^2 - 2\gamma_2{}^2 + 6\gamma_3{}^2\right)k^4 + \left(6\gamma_2{}^2 - 6\gamma_3{}^2\right)\mathbf{T}_4{}^{abcd}k^a k^b k^c k^d\right\}$$

At this point we essentially have the energies, for if

$a = 1$

$b = \frac{1}{2}\text{spur}\left(\mathbf{H}_L\right)$ and

$c = \frac{1}{8}\left[\text{spur}\left(\mathbf{H}_L\right)^2 - 2\text{spur}\left(\mathbf{H}_L{}^2\right)\right]$ then the energies E obey

$aE^2 + bE + c = 0$

The energies obtained agree with the energies known by other methods. We have repeated the above, with stress terms added to the Hamiltonian. Most limiting cases can be done fairly easily "by hand," but the case of stress in an arbitrary direction, and without assuming that stress terms dominate over kinetic terms, is quite complex.

At this time, the computer manipulation of the Luttinger and related Hamiltonians shows promise of producing intermediate results that are simple enough to be comprehensible to human physicists, and these methods are being used to compare the various formalisms for band-theory calculations (i.e., Luttinger, vs. 8-band, vs. 6-band, etc.). The original motivation for coding the simulator was as an aid to estimating scattering processes within devices. It is expected that such work will proceed, although there was not time to pursue it extensively under the present contract.

It should be added that the symbolic manipulator has a few quirks. For example, it fails to recognize that

$$0 = \varepsilon^{\mu\nu\rho}\mathbf{J}_3^{\mu\alpha\beta}\mathbf{S}^{\nu\alpha}\mathbf{S}^{\rho\beta}$$

where $\mathbf{S}$, and $\mathbf{J}_3$ are symmetric tensors (it is told the symmetry property of all symmetric tensors). Fortunately, such combinations have not arisen often and can be "fixed" by hand, and, in any case, they do not make the result "incorrect"

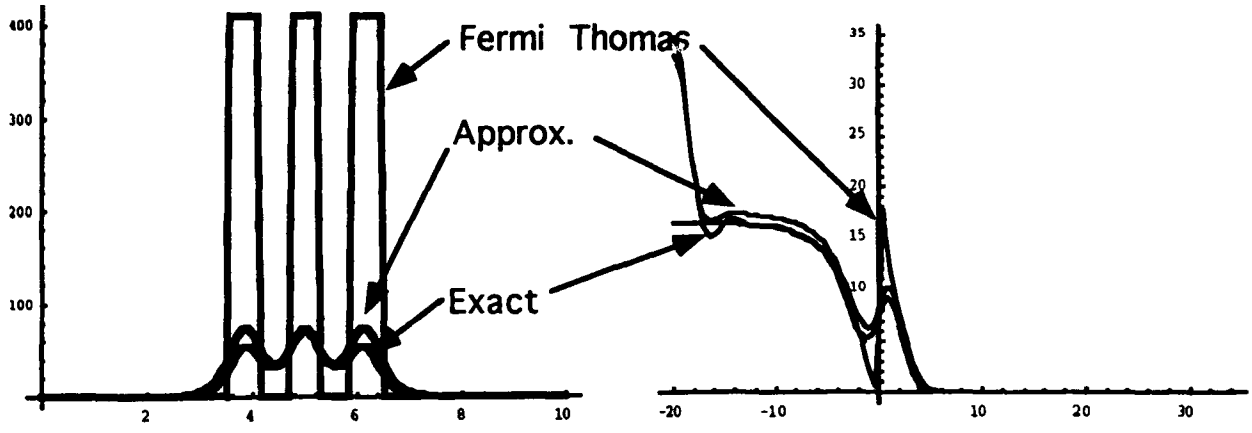
Better Approximation to Quantum Charge

One interesting byproduct of our research has been an expression for the charge in the Hartree approximation, which incorporates quantum effects but is numerically easy to compute, as one does not have to solve an eigensystem. This formula has been

demonstrated at this time under restrictive assumptions, but the concepts used to derive the original formula can be improved upon, so that quantum-charge densities can be obtained for multiband models, and the overall accuracy can be improved as well. It should even be possible to generalize the formula to an approximation of the tunnel currents, so that numerically-accurate descriptions of the device physics become available by inverting matrices instead of solving complete eigensystems.

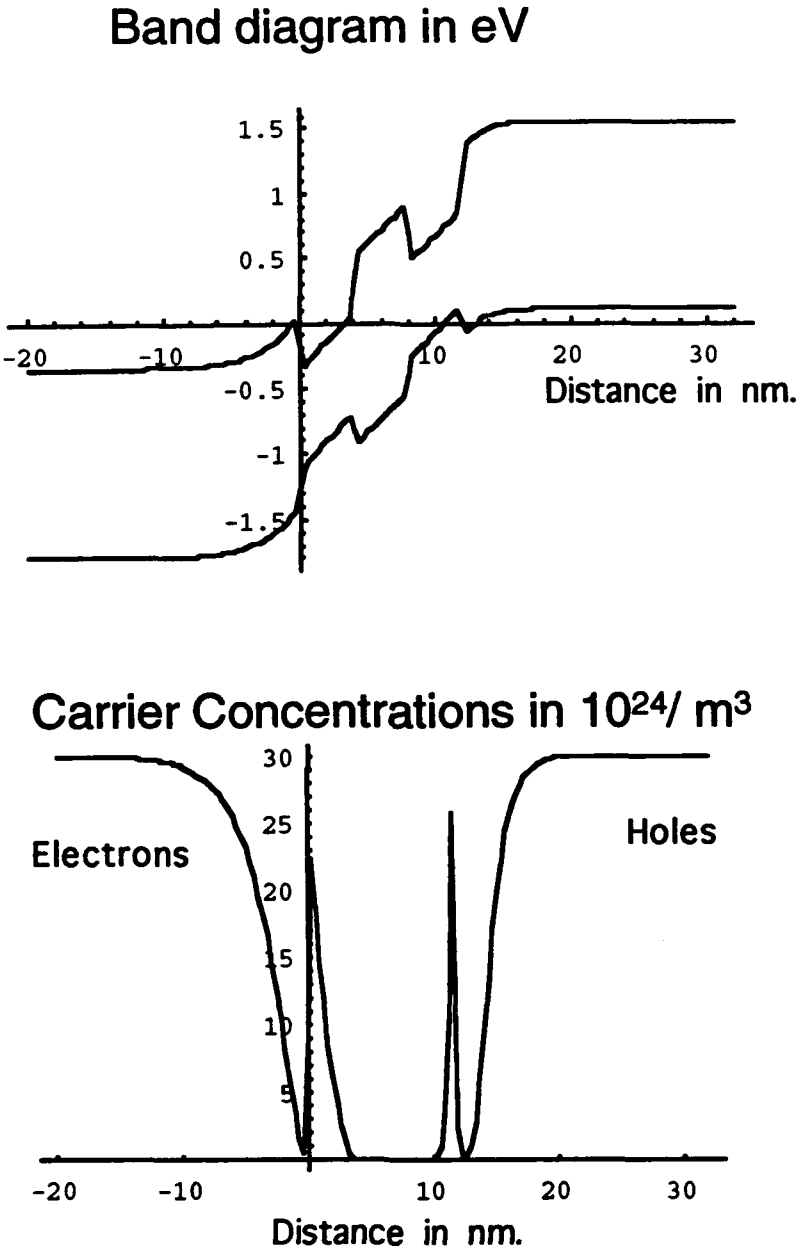
The Poisson solver in our simulator uses a Fermi Thomas expression for the charge density:

$$n = \frac{g}{(2\pi)^2}\left[\frac{2m}{\hbar^2}\right]^{3/2} \beta^{-3/2} F_{1/2}\left(\beta(\mu - E + \phi)\right)$$



This shows two examples of the quantum charge approximation. In both cases, there are three curves: one is the "exact" Hartree charge, one is the Fermi Thomas charge, and the approximate quantum charge is shown too. Figure 8a shows a set of deep multiple quantum wells, with a relatively low m. In this case, the Fermi Thomas approximation not only misses the "smearing out" effect of quantization, which is to be expected for any rapidly varying potential, but it grossly overestimates the total charge, which is worse. Figure 8b uses a "real" potential obtained using our Poisson solver to self-consistently solve Poisson's equation and the number density equations for a tunnel device. In this case, the overall agreement between the Fermi Thomas approximation and the quantum charge densities is good, indicating that the Fermi Thomas approximation can be used, but the quantum charge densities are spread out, as one would intuitively expect. The sharp rise in quantum density at the left side of the figure is a result of the fact that Neuman boundary conditions were imposed. The quantum approximation tends to overestimate the charge, but this would be easy to correct for, as the overestimation will generally be roughly proportional to the Fermi Thomas charge.

Figure 11



Band diagram and carrier concentrations for a double well device with 4nm wells and a 4 nm barrier.

Figure 12

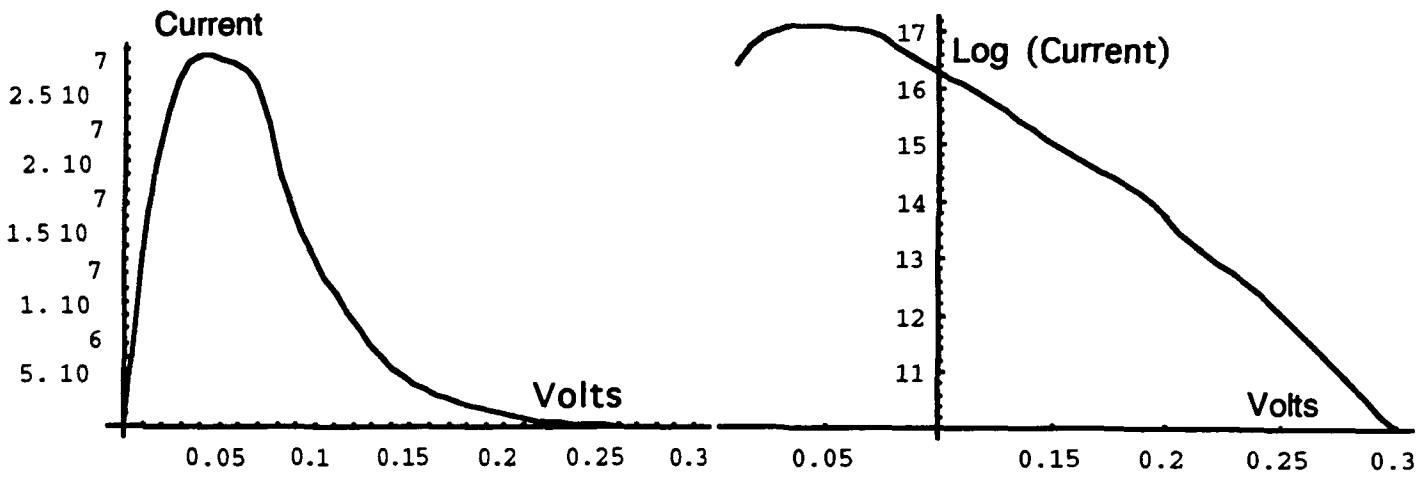


Figure 13a

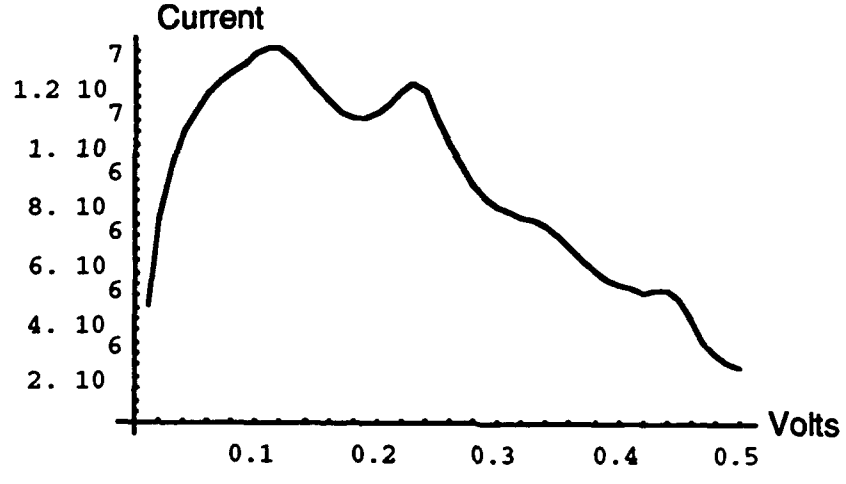
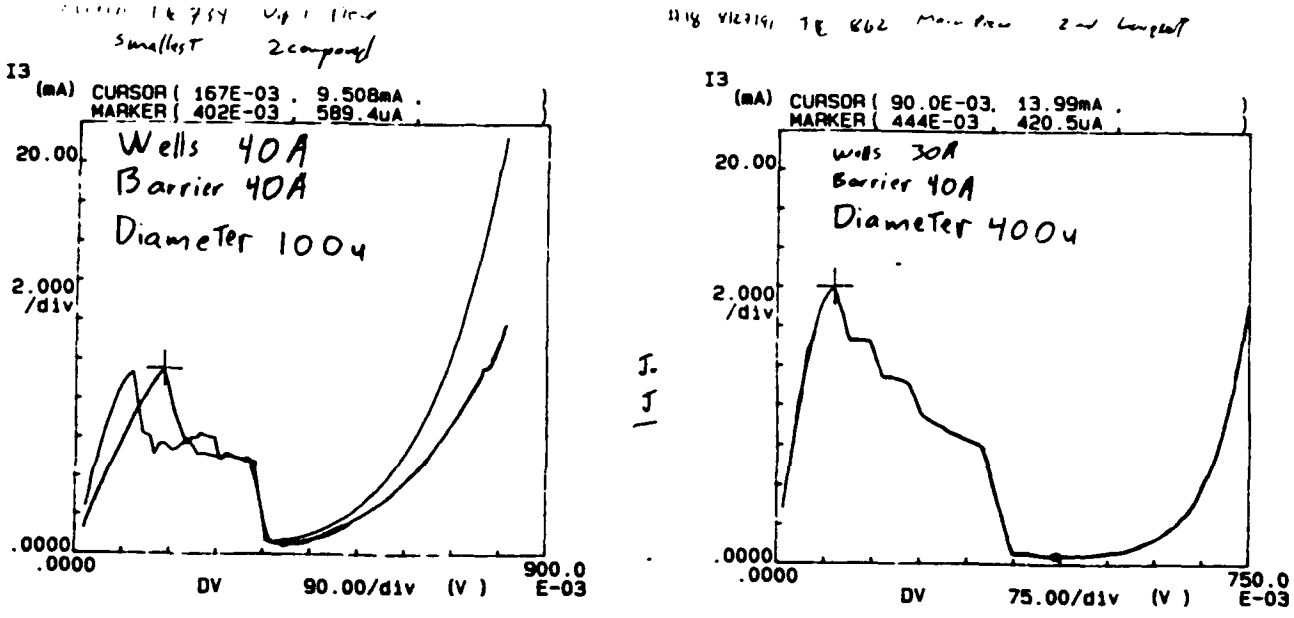


Figure 13b

Computed I/V characteristics for two devices. Fig 13a shows a double well device with 4nm wells and a 2nm barrier. The shape of the I/V curve is typical of "reasonable devices. Notice the linear slope of the log plot. Fig 13b shows a barrier only device, with GaInAs contacts and a 4nm barrier. This device had a doping level of 30 1024 on the n side and 15 1024 on the p side. The density of states mass was left at 0.048 on the n side, leading to an unphysically high Fermi energy.

Figure 13



I/V characteristics measured for RITD devices. Both are double well, single barrier RITD's. The first has 3nm wells and a 4nm barrier, while the second has both wells and the barrier 4 nm long. The 4nm well device shows two curves, which are for two different devices on the same wafer. The maximal currents are well matched, but no the series resistance. Near the edge of the wafer the maximal currents increase.

The negative resistance parts of the curves show are not stable, and the parameter analyzer shows only artifacts of oscillation at these points.

Figure 14

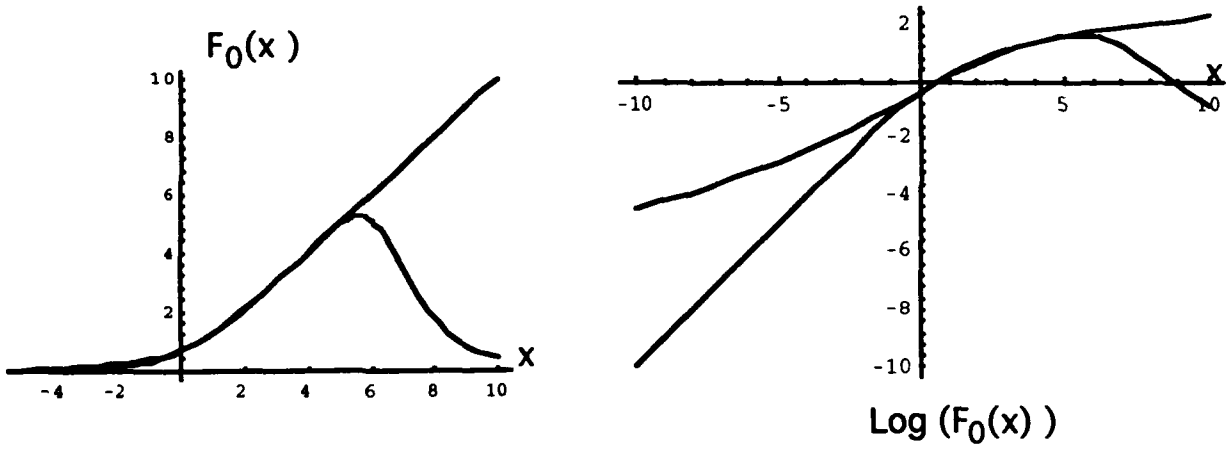


Figure 10

This compares the Fermi function $F_0(z)$ to a rational approximation. Both a linear scale and a log scale are used. On a linear scale, agreement is good for all z below about 6, at which point the approximation falls rapidly while the real Fermi function continues to grow linearly. Looking at the log scale, we can observe that, while the approximation is "small," the real Fermi function is smaller. This is to be expected, as no rational function can match exponential decay.

The Hartree approximation, on the other hand, requires us to solve the eigensystem for the perpendicular degree of freedom (1 dimensional quantum well). It then fills in the charge separately for each eigenvector:

$$n(x) = g\frac{2m}{\hbar^2}\beta^{-1}\sum_{E_k}\overline{\psi}_k(x)\psi_k(x)F_0(\mu - E_k)$$

where we have assumed, for now, that the degrees of freedom parallel to the quantum well are factored out from those perpendicular to the quantum well.

Our approximation is based upon the density matrix formulae:

$$n(x) = \rho_{xx}$$

$$\rho_{xy} = g\frac{2m}{\hbar^2}\beta^{-1}\sum_{k}\overline{\psi}_k(x)\psi_k(y)F_0\left(\beta(\mu - E_k)\right)$$

where the function $F_0(...)$ can be defined by a power series or other means. A key step is to approximate F_0 by a rational function that most of our work has used:

$$F_0(z) \approx 3.5\left[P\left(1-\frac{z}{a}\right)\right]^{-1}$$

where

$$P(x) = 1 + x + x^2 + x^3 + x^4 = \frac{1-x^5}{1-x} \quad [13]$$

Figure 10 illustrates the accuracy with which this function approximates $F_0(z)$. If the Hamiltonian has an eigenvalue, E_k, such that $\beta(\mu - E_k) > 6$, then the approximation will fail, but one can re-scale the system by a factor a according to

$$F_0(z) \approx a\left[P\left(1-\frac{z}{a}\right)\right]^{-1}$$

While this is not as good an approximation as equation 13, it will generally be fairly accurate, and in fact this re-scaling will usually be needed except at very high temperatures.

The point of all this is that evaluation of the rational approximation requires only matrix multiplication, and inverse, and not eigensystem analysis. For sparse matrices, it will often be preferable to evaluate the rational function by means of partial fractions, because multiplication will quickly fill in the zero elements of H, while a banded matrix with narrow width can be quickly inverted.

Figure 11 illustrates the actual use of this approximation for a scalar Schroedinger's equation. The agreement turns out to be excellent. Referring back to Figure 10—the rational approximation fails to fall off rapidly as $z \rightarrow -\infty$. The asymptotic distribution of eigenvalues of H is known (this is after all the basis of the Fermi Thomas approximation), and so this effect can be estimated as a simple function of potential and μ, and subtracted off. It will in fact be nearly proportional to the simple Fermi Thomas charge.

A few comments are in order. The first is that the method of using simple rational (operator) approximations for "quantum" effects is relatively novel in this context, and we are attempting here to show its promise, not its ultimate potential. The general concept has considerable flexibility, and possible extensions are:

In the language of operators, the Fermi Thomas approximation uses the trace of $H\delta(x)$ as the sole basis for estimating charge, while our proposed approximation uses the trace of operators of the form $(H - \lambda I)^{-1}\delta(x)$ as a basis. Clearly, the more information used the better, and one would expect that an optimized approximation would use both.

The methods outlined here should be applicable to the problem of estimating transport. The idea that "hydrodynamic like" equations can describe quantum transport is said to have been espoused by Madelung and Schroedinger in the 1920's. To my knowledge, this view has attracted relatively little attention. I must give some credit, however, to Dr. H Grubin, at Scientific Research Associates, who has advocated the use of "hydrodynamic like" equations for the understanding of semiconductor devices.

Simulations

All of the simulations carried out have been based, to some degree upon a set of devices fabricated at the Varian Research Center, in Palo Alto C., and the details are described in the section on those devices. Experimental variation of the materials and device parameters were carried out and some general conclusions can be drawn, especially with regard to the sensitivity of device characteristics to materials parameters. A great amount of detail of device operation was computed, Figure 12 illustrates a band diagram ,and carrier concentrations, while Figure 13 shows some computed I/V curves. The Shape of I/V curves of devices with physically "reasonable" parameters tended to all look similar in shape to Figure 13a, a single sharp peak at a very low voltage of roughly 0.05 V. Figure 13b shows a device with a Fermi energy that is unrealistically high, on the electron side. The multiple peaks in this I/V curve are not completely understood.

Real devices tended to have a maximum at a much higher voltage, and this was generally attributed to a parasitic resistance in series with the diode. I would point out that this maximal voltage tended to vary greatly from one physical device to the next, among the Varian devices. For comparison, Figure 14 shows some measured I/V curves.

Simulations Relevant to the Varian RTD Devices

The Varian research center in Paulo Alto California fabricated a large set of RITD's, and although the project was discontinued, most of the devices fabricated so far are now in the possession of the University of Toronto EE Department. I go there occasionally, and this summer rechecked the I/V characteristics of a large set of devices, which were at the University during that visit. Those devices are the ones used in this research. They were made up of 8 double-well RITD's, 2 RITD's with wells only and no barriers, and one

device with low band-gap contacts and a barrier. The double-well devices are all symmetrical—the left and right wells are of equal width. Varian has fabricated several other sets of devices, some of which have been described in publications, most notably a set of RITD's with barriers only varying in height. All the devices had contacts doped to 30 10^{24} m^{-3}, and used a well material of InGaAs, with barriers of InAlAs. Except for the barrier-only device, the contacts were also InAlAs. In all devices, the doping was in the contacts only, not (intentionally) extended into the wells or barriers. The well and barrier materials used were lattice-matched to the underlying InP substrate, so that these were unstrained devices. Table I summarizes the Devices.

Figure 15 condenses our results. Test cases with artificially-high density of states had more "scatter" than those with the theoretical density of states, even though the theoretical density resulted in an overall overestimation of the current density. The systematic agreement is very good, if devices 1, and 9 are excluded. Device 9 is clearly an anomaly, as it is physically an related to devices 2, 1, and 5 which show a greatly increasing current with decreasing barrier width. Device 1, which is not so far off, in any event, is an "anomaly" in the other direction. Devices with parameters of 1 had the best performance of all the Varian RITD's, they had high current densities and very high peak to valley ratio's.

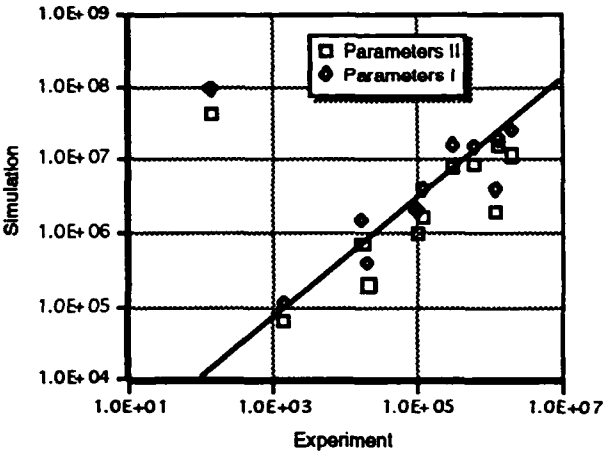


Figure 15

The comparison of theory and experiment. This shows maximal currents for all the Varian Devices and their simulated values. The graphs shows two parameter sets, parameters I set the densities of states masses to 0.1 an d 0.083 for the conduction bands in InAlAs, and GaInAs, respectively. Parameter set II sets the masses to 0.122 and 0.1. Parameter set I is modeled after the physically "best" value, and can be fitted closely to the experimental data, while parameter set II is closer to the absolute value of the data. The slope of the line is approximately 0.78

Device and Materials Parameters

Table II summarizes the materials parameters used as our "standard" set.

Table I lists the Varian devices, including all devices that were readily available for a recheck of the I/V characteristic when I was in Toronto in the summer of 1991. Some devices did not show negative resistance at all and were not included.

Table I

This is a table of the Varian devices used in this study. The material and doping parameters are described in the text, and in Table II.

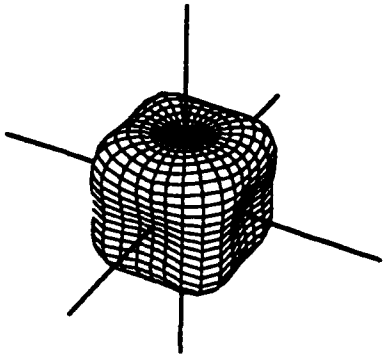
Varian Resonant Interband Tunnel Devices

Double Well Devices-----both wells equal			
Device	Well nm	Barrier nm	Maximal Current A/m^2
1	4	4	$1.27\ 10^6$
2	4	6	$2\ 10^4$
3	6	4	$1.75\ 10^4$
4	6	2	$3\ 10^5$
5	4	2	$2\ 10^6$
6	3	4	$1.2\ 10^5$
7	2	4	$1\ 10^5$
8	6	6	$1.4\ 10^3$
Devices with Well only			
9	8	------	139
10	4	------	$6\ 10^5$
Devices with InGaAs Contacts, and a Barrier only			
11	--------	4	$1.3\ 10^6$

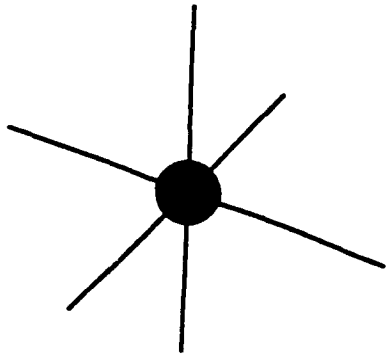
Table II

This describes the materials parameters used for simulations. In some cases specific parameters were varied, as described in the text. All devices were considered to be composed of two materials, a barrier material (AlInAs), which was generally used as the contact too and a well material (GaInAs).

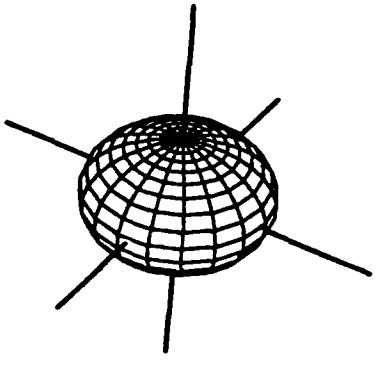
Parameter	Barrier/Contact	Well	Description
E_c	1.4 eV	1.0eV	Conduction Band Level
E_v	0 eV	0.25eV	Valence Band Level
M_v	0.52 m_e	0.52 m_e	Valence band Mass --- This is a density of states mass used by the Poisson solver to compute charge density
M_c	0.1 m_e	0.083 m_e	Conduction band Mass --- This is a density of states mass used by the Poisson solver to compute charge density
χ	6.5 eV	6.5 eV	Band Parameter: $\chi = \frac{2m_e}{3\hbar^2}\lvert P\rvert^2$ P is the interband matrix element
m_c	.086 m_e	.048 m_e	Conduction Band Mass used for solution of Schroedinger's Equation
m_v	.091 m_e	.052 m_e	Valence Band Mass used for solution of Schroedinger's Equation
m_{pv}	0.3 m_e	0.3 m_e	"Parallel" valence band masses, used to adjust the well energies as parallel momentum is increased
m_{pc}	0.86 m_e	0.48 m_e	"Parallel" Conduction band masses, used to adjust the well energies as parallel momentum is increased---These are the same as the normal masses.



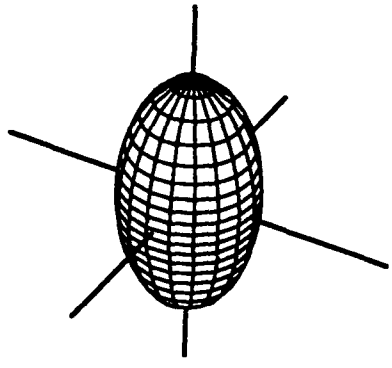
Heavy Holes No Stress (a)



Light Holes No Stress (b)



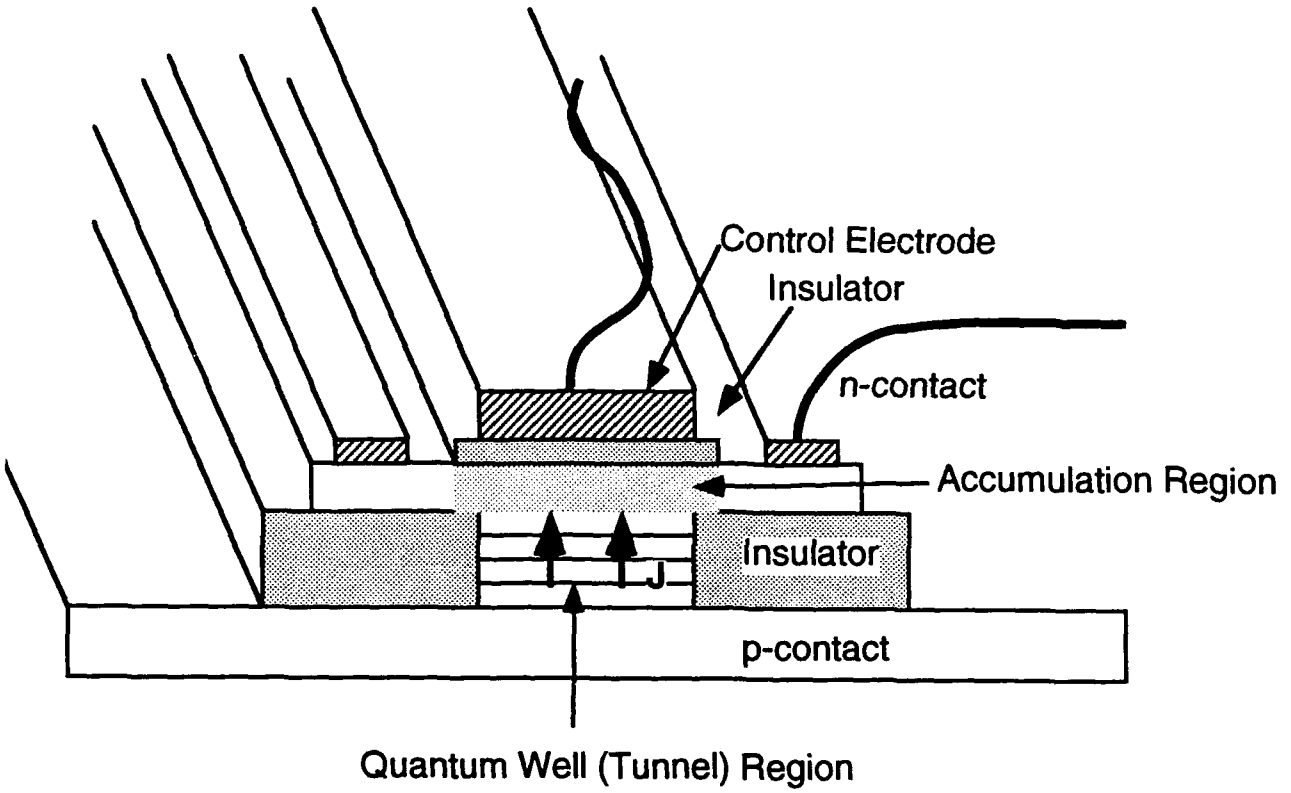
Hole Band with Stress (c)



Hole Band with Stress (d)

This illustrates the effect of stress upon the valence bands of a semiconductor. The band parameters are those of GaAs, but are believed to be typical of a high-performance semiconductor. The graphic shows a constant energy surface in momentum space. In the absence of stress, the heavy holes have constant energy surfaces which resemble rounded-off cubes. The shape reflects the cubic symmetry of the lattice, while manifesting large anisotropy. In 13b, the light holes are shown on the same scale, and one can observe that the volume is dramatically smaller. On this scale, it is hard to see the light hole surface, but when expanded, it does not depart visibly from a sphere.
Parts c and d show the two bands near G after the application of a large stress. Although the shape is different, the two ellipsoids have approximately the same volume. The lines in c and d are 1/2 the length of the corresponding lines in Figure 13a and b, with the stress in the z direction—the direction of the "longitude" lines. For compressive stress (squeezed in the z direction, pulled in the x and y directions) 13c will be at the top of the band.

Figure 16



This shows the "basic" 3-terminal device structure. A complimentary device exists, with the n and p regions interchanged. The action of the control electrode is both to influence the shape of the n quantum well and to change the n carrier concentration. A real device might be long in the z direction (into the paper) while taking the form of a narrow strip, as shown, to reduce parasitic resistance. The barrier below the control should be 25-30 nm to reduce parasitic conductance, and while the carrier concentrations in the n well should fill the first sub-band, be on the order of $2\ 10^{16}\ m^{-2}$. The control voltage would be on the order of 1 volt, so that the insulating layer might have to be a "real" insulator, not a high band-gap semiconductor, in order to limit leakage currents.

Figure 17

Effect of Stress

Stress is of interest for many reasons, ranging from the desire to combine materials of differing lattice constants, in which case the stress may be incidental, to the hope that stressed devices will perform better. Reference 18 is an example of a relatively novel tunnel device which combines materials of differing lattice constants and obtains very high performance. While it is not "interband" in the sense uses in this study, it does involve tunneling from the Γ valley to the X valley. It also demonstrates a device technology compatible with the "mature" GaAs material system.

As discussed earlier, stress has a large effect upon the valence band structure. This is illustrated in Figure 16, which shows a set of 3 dimensional constant energy surfaces in momentum space. In the absence of stress, the heavy holes have a complex energy surface, which is more nearly cubic than spheroidal, while the light hole energy surfaces are very nearly perfect spheres. After the application of stress, the bands become intricately mixed, and the result is two new bands separated at the Γ point, and with spheroidal energy surfaces. In the case of GaAs, the two spheroids have almost identical volumes, and the resulting effective (density of states) mass is 0.17.

The dominant effect of stress upon an RITD is expected to be the, dramatically decrease the density of states and this is the only effect which can be readily incorporated into the two band model. We decreased the density of states mass to $0.18m_e$ for the valence bands of stressed devices. This value was obtained by calculating the density of states of the Luttinger model, as discussed earlier. Several Simulation runs were made assuming that a piece of the quantum-well structure was stressed by compression in the z direction. If the doping were left at the "standard" 30 10^{24}, and the density of states for the quantum wells was changed to 0.18—the number indicated by the band density of states before the structure becomes too nonparabolic—then the maximal current of a 4 nm well 4 nm barrier double well RITD doubled. If a 3 nm region next to the well had a reduced density of states, the current increased by a mere 25%. If both the well and a 3 nm region were stressed, then the increase was roughly 250%. Altogether, the main effect would seem to be due to a reduction in the charge stored in the well under stress, which can double the critical current.

Other Parameters

Sensitivity of the barrier heights and widths was very strong, as is to be expected on the basis of the WKB approximation. In addition,the maximal current was most sensitive to parameters which changed the Fermi energies relative to the band edge, When graded doping was put into the well without decreasing the doping outside the well, then the current increased over 20. On the other hand, if the doping was graded, but with no net increase (ie taken from the contact and put into the barrier) then there was no significant change in maximal current.

The results were not sensitive to a change in band offset, *or to the band parameter* χ.

Conclusions

Computer simulation using a 2-band model predicts the trends well for Interband tunnel devices, but the absolute values are off by a systematic amount. It is not known why the simulated results show a dramatically larger current than is observed, but it is thought to be related to the effects of high doping in our test "population".

Compressive stress in the quantum well on the p side, and doping in the barrier, are both predicted to increase the maximal current.

Although the simulator did not add in the thermionic current, analytic estimates of this current were discussed in an earlier section, and they do not account for the observed valley currents, which are higher.

Three-Terminal Devices

There appears not to be a satisfactory means to contact the interior of an RITD, so the primary 3-terminal device structure available involves the addition of an electrode outside the device, as shown in Figure 17. At this time, we have only begun to explore the potential of such devices, but it would appear that it is possible to modulate the tunnel current by means of a control voltage applied to the control electrode. This device structure is very similar to the Stark Effect transistor, and the control voltage would affect the quantum well, but in an RITD, more transconductance is likely to arise as a result of modulating the carrier concentrations in the well.

An estimate of performance can be made as follows assuming a "high"performance current density:

Current with one volt applied	10^7 A m^2;
Capacitance	3 10^{-3} F m^{-2}

The capacitance is just (ε/d).

In short, one can modulate the RITD, but the transconductance does not yield a device which can compete head on with FET's for applications in logic and signal amplification. On the other hand, if one is using the RITD as a comparison with a variable threshold, or as an oscillator, then such a control electrode would be very valuable.

Of course, many device physics experiments could be carried out upon such structures as part of a basic research program.

Physics Issues

When we first began to obtain simulation results that we felt were reliable (in the sense that the numbers represented the model, not necessarily reality), it was immediately observed that the tunnel currents were predicted to be much (10-100 times) larger than observed. We naturally questioned what the discrepancy was due to.

8-Band Model

It is widely agreed that the 8-band Kane model should be an accurate representation of the band structure near the Γ point. The 2-band model used in our present simulator is only an approximation to this. Nonetheless, it is expected to capture most of the physics of the interband mixing. McLellan *et. al.* carried out simulation (of a different type of device) and compared the 8-band results to 2-band results. They confirmed that the 2-band results are usually within a factor or two of the 8-band results, and very rarely more than a factor of 10 off. They did not report a systematic deviation between the two models, as we have found between our model and the devices fabricated at Varian. There is some question as to what 2-band parameters are relevant to a given set of 8-band parameters.

Band Parameters

There is some uncertainty in the band parameters used in our simulations. This uncertainty has two sources. First, some of the band parameters, especially χ, are not well known. It appears, however, that the tunnel rate is relatively insensitive to χ. Second, the 2-band model parameters to be used are not always related to the 8-band parameters. Finding the optimal correspondence between 8-band and 2-band parameters could be the subject of future work.

Heavy Doping

The Varian devices used extremely high doping, and high doping is often a prerequisite for a high current density (i.e., fast) device. In out opinion, the systematic deviation of the measured devices relative to the predictions of our simulations is most likely due to the fact that tunnelling into and out of highly doped regions involves physical effects which we did not model correctly. The relevant physics itself is not well understood, and what is needed is more basic research on highly doped materials.

Charge Density

Our simulator used a simple Fermi Thomas charge density, assuming parabolic energy bands. Analytic estimates showed that this is considerably far off, at the doping densities used, and an attempt was made to "fix" the problem by utilizing a higher density-of-states effective mass. This is not wholly satisfactory, for the actual carrier densities vary from near 0 to 30 $10^{24}m^{-3}$ in the Varian devices, and our effective mass, which was chosen to result in the correct density at 30 $10^{24}m^{-3}$ will overestimate the density as μ-E_C becomes lower. While it should not be difficult to incorporate a more general charge density calculation into a simulation program, there is the possibility that at the highest densities carrier interactions will provide additional effects, which will affect device performance.

We estimated the difference between the Hartree quantum charge density and the Fermi Thomas charge density, and found it to be small in the Varian devices. The departure from parabolic bands is a much larger effect.

Thermionic Emission

The thermionic emission is given by the same formula used to calculate interband tunnelling:

$$J = \frac{gq}{2\pi\hbar}\int\frac{d^2k_p}{(2\pi)^2}\int dE|T|^2(f_L - f_R)$$

The thermionic emission is simpler, for several reasons:

The Fermi factors can be replaced by Boltzman statistics,

$|T|^2$ is almost exactly 1 in the classically allowed region. This is not just an assumption, our simulator has been used to calculate $|T|^2$ and it rises from 0 to 0.8 over a range of 6meV as the energy rises above the conduction band on the right.

In the end, one obtains the well known expression:

$$J = \frac{gqm}{(2\pi)^2\hbar^3}(K_BT)^2 e^{(V-V_T)q/K_BT}$$

where T is now temperature not transmission, and V_T is the threshold, and where the left and right conduction bands are lined up.

This formula predicts that J will be below 10^4 Am^{-2} 0.4 Volt below the gap, or ~ 1 Volt for InAlAs. Most of the devices show a large rise well before this, and the cause may be "indirect" conduction between the wells, as discussed earlier.

Scattering

Figure 1, especially Figure 1c, emphasizes that scattering can be a dominant means of conduction, especially when the bias voltage is in the "gap" just below the region of direct thermionic current. In other cases, too, scattering can affect the I/V characteristic and other manifestations of device performance. The present work did not take scattering into account in any quantitative manner, and this should be changed in future work.

Applications Issues

This short section will necessarily have the flavor of personal opinion rather than hard scientific fact, but with this warning you may read on. Most of the comments are fairly generic to hysteretic tunnel devices, Josephson Junctions, Tunnel Diodes, Resonant Tunnel Diodes, and Interband Tunnel Diodes. Tunnel devices have many potential applications in the generation of very high frequencies, and in the generation of sharp step waveforms. The upper frequency limit is essentially given by the band gap, and tunnel devices can, in principal, provide power in the far infrared, where existing technology is very limited. Tunnel diodes are used to generate fast-rising pulses for instruments, especially sampling

oscilloscopes, and time-domain reflectometers, and such applications are likely to continue. The fast responses of which a tunnel diode is capable are well suited to the comparison needed for some A to D applications, and as a pulse discriminator in digital communications. The ability to control the threshold with a "3rd" terminal would be valuable in these applications. At the risk of sounding negative, however, I would discourage expectations that these devices are likely to displace transistors for general-purpose logic, or signal amplification.

Past research upon two terminal devices has tended to focus too heavily upon the possible applications as logic gates. The IBM Josephson computer project, which ended in the early 1980's, is an example of this. It has gradually been learned that tunnel devices are not well-suited to general-purpose logic. The reasons for this cannot be elucidated in this report, but the core problem is that circuits have poor margins. The ability to trade off gain for margins is not available for low-gain devices, and the matter is made worse when the current density depends exponentially upon device-fabrication parameters.

Similarly, the elimination of harmonic distortion by means of negative feedback is only available for high-gain devices, so that negative resistance amplifiers have often been plagued by harmonic distortion and limited dynamic range.

High gain without instability is inherently difficult to achieve in a two-terminal device without any isolation between input and output.

"Three-Terminal Device" has come to be a Holy Grail in the Superconducting micro-electronics industry, and reflects a desire to find the way to attain the high speed and low power dissipation of tunnel devices, with the isolation and gain which is more typical of today's highly developed transistors. At this time, Heterostructure Bipolar Transistors routinely exhibit current gains of 5,000, and can provide useful gains in the mm wave region of the spectrum. The prospects for finding tunnel devices which are a quantum improvement over this (or MODFET's etc.) are not good, and one is well advised to

Recommendations for Future Research

This project has led to several recommendations. In this project we have begun the development of software tools for device physics, which can have an impact upon the field.

Software Tools

During this work, the ability to interactively explore device physics was very rewarding. Our simulator will generate a band diagram in a few seconds, and future simulations should attempt to provide such rapid feedback, even at the expense of sophisticated models.

Our user interface worked far better than other simulator interfaces used by this author, and although means should be found to make the programming easier, it too is clearly the direction in which to proceed.

In the end, device simulation tools, should be designed for use by "non-experts", so that, for example, the researcher in a fabrication group can have access to state of the art simulations based upon the devices under fabrication.

Characterization Tools

The primary tool for the characterization of two terminal devices is the I/V curve extractor. A core problem is the desire to obtain full I/V curves of devices with a negative resistance region. Most researchers use a "parameter analyzer," usually the Hewlett Packard product, which is optimized for very low leakage current, but is rarely stable when taking the I/V of an interband tunnel device. An ideal negative-resistance device will be stable when connected to a low impedance, as illustrated in Figure 18. Devices with low current densities (and thus slow switching speeds) can be "stabilized" by adding external circuitry, but this becomes more difficult for high-current-density devices, as parasitic inductance in the

interconnection can cause a device to oscillate internally even when it is "connected" to a low impedance, because the physical device sees the packaging inductance. At this time, relatively few laboratories are performing high-speed testing upon tunnel devices.

General-purpose "I/V curve" measurement electronics are usually "home-built" because the main manufacturers have not made parameter analyzers suitable for use with conditionally stable devices. It is especially desirable to have a means to extract the intrinsic I/V curve from a device which cannot be stabilized at all. In principle, this could be done by analyzing the harmonics generated when the device is swept with a large amplitude (microwave) sine wave, so that the device switches exactly once each cycle. To extract an I/V curve from such, or similar, data requires software that is not generally available.

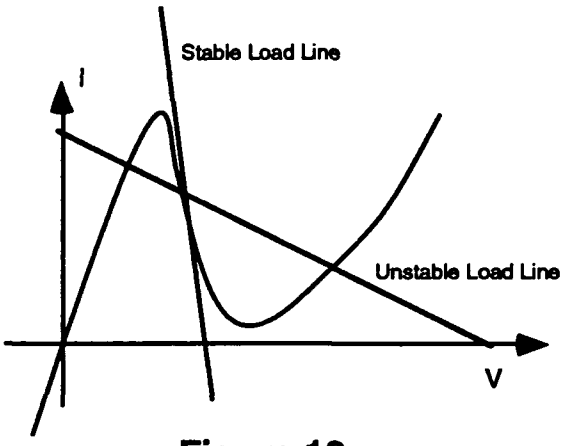


Figure 18

This illustrates a "generic" tunnel device I/V curve, and shows both a stable load line, and an unstable one.

Of course, these devices are often intended for use in very high speed circuitry, and so it is desirable to be able to perform high speed testing. This opens up special problems as present high speed testing centers upon the use of microwave network analyzers for characterization of linear devices. The characterization of nonlinear devices is an area where improved test equipment is badly needed.

Three-Terminal Devices

Research upon three-terminal devices should consist of two parts. As interband tunnel devices are presently pushing the frontiers of device physics, it is useful to view some devices as valuable research tools, without regard for ultimate application. This is not to say that applications will never be found for Stark effect transistors, for example. The near-term applications are most likely to involve using a control electrode to modify the behavior of a device that is used essentially as a two- terminal device. At this time, three-terminal devices with sufficient gain to be useful for general-purpose logic or signal amplification appear unlikely.

One dimensional Quantum Well Wires (QWW) offer the possibility of three terminal tunnel devices. This project did not permit detailed simulations of QWW devices, but our understanding of the basic physics is not fully adequate, in any case. QWW tunnel devices are however a broad and promising area for future research.

Electro-Optic Devices

Resonant Interband "Tunnel" devices can be expected to respond to light, or to generate light. In fact, the boundary between "optical" and "electrical" can become blurred within such a device. The generation of sub-mm waves is an area where interband tunnel devices are promising, and competing technologies do not work well. The fundamental frequency limits are set by the bandgaps (0.75 eV corresponds to 181TerraHz. In practice, parasitic losses will limit performance. The practical limits are not known, but efforts should be made to design device structures for low RF losses.

It would take us too far afield, to discuss the theory of photon assisted tunneling in detail here. The theory of photon assisted tunnelling in resonant tunnel devices has been discussed by this author[19] A detailed theory of quantum

effects in tunnel devices was published by Tucker[20], and is bases, as was our own paper, upon the transfer matrix approach of Bardeen[21] and others[22].

References

[1]"Resonant Interband Tunnel Diodes", M. Sweeny and Jingming Xu, Applied Physics Letters **54,** 546, February, 1989

[2]L. F. Luo, R. Beresford, and W. I. Wang, "Interband Tunnelling in Polytype GaSB/AlSb/InAs heterostructures", *Appl. Phys. Lett.*, vol 55, pp 2023-25 1989

[3]J. R. Soderstrom, D. H. Chow, and T. C. McGill, "A new negative differential device based on resonant interband tunneling"", *Appl. Phys. Lett.*, vol 55, p 1094 1989

[4]L. Esaki, "New Phenomena in narrow germanium p-n junctions", *Phys. Rev.*, vol 109, p. 603, 1958

[5]J. M. Xu, Private communication, One of the Varian devices which has only recently been tested has broken the peak to valley ration of 100 "barrier". This was a device near the edge of the wafer, but evin "typical" devices had high ratio's as can be seen from figure 17, of this report.

[6]D. Z. Y. Ting, D. A. Collins, E. T. Yu, D. H. Chow, and T. C. McGill, Appl. Phys. Lett. **57,** 1257, (1990)

[7]"Multiband Treatment of quantum transport in interband tunnel devices" Z. Y. Ting, E. T. Yu, and T. C. McGill, Phys, Rev B vol 45, no. 7, 3583 15-Feb-92

[8]"Interband Tunneling in Heterostructure Tunnel Diodes", R. Q. Yang, M. Sweeny, D. Day, and J. M. Xu, IEEE Transactions of Electron Devices, vol 38, no. 3, Marh 1991

[9]"Double Quantum well resonant tunnel diodes", D. J. Day, Y. Chung, C. Webb J. N. Eckstein , J. J. Xu, and M. Sweeny, Appl. Phys. Lett., vol. 57, No. 12, pg. 1260, 17 Sept.1990

[10]W. Shockley, Phys. Rev. 78, 173 (1950)

[11]"Cyclotron Resonance of Electrons and Holes in Silicon and Germanium Crystals", G. Dresselhaus, A. F. Kip, and C. Kittel Phys. Rev. vol 98, no. 2 pp 368-84, 15-Apr-1955

[12]E. O. Kane, J. Phys. Chem. Solids 1, 245 (1956)

[13]E. O. Kane, J. Phys. Chem. Solids 12, 181 (1959), Keldish, and others have used this equation, at roughly the same time.

[14]"Theory of Brillouin Zones and Symmetry Properties of Wave Functions in Crystals", L. P. Bouchaert, R. Smoluchowski, and E. Wigner, Phys Rev vol 50 pp 58-67, 1-JUL-1936

[15]J.M. Luttinger and W. Kohn, Phys. Rev. 97, 869, (1955).

[16]"Numerical Recipes in C" W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T Vetterling, Cambridge University Press, Cambridge, UK, 1988,

[17]"Inside Macintosh" Vol I - VI, by various authors at Apple Computer Inc. Addison-Wesley, Reading Mass. 1985

[18]"Highly strained GaAs/InGaAs/AsAs resonant tunneling diodes with simultaneously high peak current densities and peak to valley ratios at room temperature", R. M. Kapre, A. Madhukar, and S. Guha, Appl. Phys Lett, **58** (20), 20 May 1991

[19]"On Photon-Assisted Tunnelling in Quantum Well Structures", M. Sweeny and Jingming Xu, IEEE Journal of Quantum Electronics vol. 25, no. 5 May, 1989.

[20]"Quantum Limited Detection in Tunnel Junction Mixers", J. R. Tucker, IEEE J Quantum electronics, Vol QE-15 no 11 Nov 1979

[21]"Tunneling from a Many particle point of view", J. Bardeen, Phys. Rev. Lett. vol 6, pp 57-9 Jan 1961

[22]"Superconductive Tunneling", M. H. Cohen, L. M. Falicov, J. C. Phillips, Phys. Rev. Lett. vol 8, pp 316-18, Apr 1962